11

Multilevel Modeling and Cross-Cultural Research

JOHN B. NEZLEK

Cross-cultural psychologists, and other scholars who are interested in the joint effects of cultural and individual-level constructs, often collect data and are interested in hypotheses that involve multiple levels of analysis simultaneously. For example, in cross-cultural research, it is not uncommon to collect data from numerous individuals in numerous countries (or cultures).[1] Such data structures are frequently referred to as *multilevel* or *hierarchically nested*, or simply *nested* data structures because observations at one level of analysis (e.g., individuals) are nested within observations at another (e.g., culture). Within a multilevel framework, questions of interest could be couched in terms of cultural differences in means of individual-level measures such as *Life Satisfaction*, within-culture relationships between individual-level measures such as *Life Satisfaction* and *Individualism*, and between-cultural differences in such within-culture relationships.

When analyzing such nested data structures, the possibility that relationships among constructs can vary across levels of analysis must be taken into account. That is, relationships between two variables at the between-country level (e.g., relationships among country-level aggregates, sometimes referred to as *ecological correlations*) may or may not be the same as the relationships between these two variables within countries (e.g., individual-level correlations). In fact, relationships at the two levels of analysis are mathematically independent (e.g., Nezlek, 2001), and it is inappropriate to draw conclusions about within-culture relationships from between-culture analyses. This inappropriateness is highlighted by the possibility that within-country (i.e., individual-level) relationships may vary across countries, undermining the validity of any estimate of "the"

[1] In this chapter, the terms *country* and *culture* are used interchangeably to denote a meaningful unit of analysis. For modeling purposes, the distinctions between country and culture are unimportant, although such distinctions can be critical substantively.

individual-level relationship, simply because there may not be a single, uniform individual-level relationship.

Some of these possibilities are illustrated in Tables 11.1 through 11.4. Each of these examples assumes a study in which *Life Satisfaction* and *Individualism* were measured for five people in each of three countries. For the data in Table 11.1, the relationship between *Satisfaction* and *Individualism* is positive in each of the three countries. As one goes up, the other goes up. In contrast, the relationship between country-level means is negative. The higher the mean score on *Satisfaction*, the lower the mean score on *Individualism*. In response to the question, "What is the relationship between *Satisfaction* and *Individualism*?" the answer should be: "It depends on the level of analysis in which you are interested." Either positive or negative could be correct.

The other tables demonstrate other possibilities. In Table 11.2, there is a negative relationship between *Satisfaction* and *Individualism* in each of the three countries, whereas at the country level, the relationship is positive. In Table 11.3, there is no relationship in any of the countries because there is no variance for *Individualism* within each country, but the relationship among country-level aggregates is positive. Finally, in Table 11.4, the relationship between *Satisfaction* and *Individualism* is different for all three countries (one positive, one no relationship, and one negative), although the relationship among country-level aggregates is positive.

Obviously, these examples do not exhaust the possible combinations of within- and between-country relationships. The point in presenting them is to illustrate that relationships at different levels of analysis are independent – any type of relationship can exist at one level of analysis simultaneously with any type of relationship at another. Moreover, as shown in Table 11.4, it is possible that within-country relationships vary across countries.

A consensus has emerged that such multilevel, nested data structures should be analyzed with what are referred to as *multilevel random coefficient models* (MRCM), sometimes referred to as MLM (multilevel modeling). In this chapter, I provide a brief review of analytic strategies for multilevel analyses, including a rationale for using MRCM and a brief discussion of the shortcomings of other approaches, typically some type of ordinary least squares analysis (OLS). I also provide guidelines for conducting and interpreting MRCM analyses, for reporting results, and for designing studies, as well as a discussion of the limitations of MRCM. That is, there may be occasions when it is conceptually desirable to use MRCM, but it is not practical or appropriate given the data at hand. General introductions to MRCM can be found in Kreft and de Leeuw (1998), Raudenbush and Bryk (2002), and Snijders and Bosker (1999).

Table 11.1. *Relationships: Positive at within-country level and negative at between-country level*

	Country 1		Country 2		Country 3	
	Satisfaction	Individual	Satisfaction	Individual	Satisfaction	Individual
	6	11	9	9	11	6
	7	12	10	10	12	7
	8	13	11	11	13	8
	9	14	12	12	14	9
	10	15	13	13	15	10
Mean	8	13	11	11	13	8

Table 11.2. *Relationships: Negative at within-country level and positive at between-country level*

	Country 1		Country 2		Country 3	
	Satisfaction	Individual	Satisfaction	Individual	Satisfaction	Individual
	1	8	9	9	9	18
	2	7	10	10	10	17
	3	6	11	11	11	16
	4	5	12	12	12	15
	5	4	13	13	13	14
Mean	3	6	11	11	11	16

Table 11.3. *Relationships: None at within-country level and positive at between-country level*

	Country 1		Country 2		Country 3	
	Satisfaction	Individual	Satisfaction	Individual	Satisfaction	Individual
	1	8	4	10	9	15
	2	8	5	10	10	15
	3	8	6	10	11	15
	4	8	7	10	12	15
	5	8	8	10	13	15
Mean	3	8	6	10	11	15

Table 11.4. *Relationships: Varying at within-country level and positive at between-country level*

	Country 1		Country 2		Country 3	
	Satisfaction	Individual	Satisfaction	Individual	Satisfaction	Individual
	1	10	4	10	9	13
	2	9	5	10	10	14
	3	8	6	10	11	15
	4	7	7	10	12	16
	5	6	8	10	13	17
Mean	3	8	6	10	11	15

ORDINARY LEAST SQUARES ANALYSES OF MULTILEVEL DATA STRUCTURES

Multilevel data structures have been analyzed with various types of analyses other than MRCM, and before describing MRCM, it will be useful to discuss these analyses and their shortcomings. Such analyses generally fall into one of two categories, *aggregation* and *disaggregation* procedures, and they generally rely on some type of OLS analysis.

Aggregation analyses were described briefly in the previous section. In aggregation analyses of a cross-cultural data set, country-level summary statistics are calculated and then analyzed. For example, mean *Life Satisfaction* and mean *Individualism* are calculated for each country and then correlated (ecological correlations). Conceptually, as discussed earlier, relationships between such aggregates cannot be assumed to describe relationships at the person level. Relationships between such aggregates are mathematically independent of relationships at the person level, and person-level relationships may vary across countries.

Moreover, unless adjustments are made to reflect differences across samples in terms of size and consistency, means based on aggregates may vary in terms of their reliability. For example, means based on larger samples will be tend to be more accurate (representative) than means based on smaller samples. Similarly, the representativeness of a mean can vary as a function of the variance of a set of observations. For example, assume two sets of five observations each: 1, 1, 5, 9, 9 and 4, 4, 5, 6, 6. For both sets of observations, the mean is 5, but clearly 5 represents the second set of observations better than it does the first set. As explained subsequently, such factors are taken into account in MRCM analyses.[2]

In disaggregation analyses, analyses are performed at the level of the individual person, usually with some type of OLS regression. In such analyses, country-level variables are often included in individual-level analyses. For example, to examine relationships between a country-level variable (e.g., *Modernization* perhaps defined in terms of the state of a country's infrastructure) and an individual-level variable (e.g., *Life Satisfaction*), the individuals in each country are assigned the modernization score for their country, and a correlation between *Life Satisfaction* and *Modernization* is

[2] It is important to keep in mind that for constructs that exist primarily or solely at the country level, such as form of government and geographic characteristics, analyses at the country level are perfectly appropriate. There are observations at only one level of analysis – the country – and so multilevel analyses are not appropriate.

calculated. Such analyses are fundamentally flawed because there can be no purely individual-level relationship between *Modernization* and *Life Satisfaction.* By definition, *Modernization* exists at only the country level, and any analysis that estimates a relationship between *Modernization* and *Life Satisfaction* that does not take this into account confounds the two levels of analysis – for example, estimates of individual-level relationships – are confounded by country-level differences. An apparent solution to this problem is to calculate a mean *Life Satisfaction* score for each country and then correlate these means with *Modernization.* As noted earlier, however, such analyses cannot be used to estimate individual-level relationships.

Relationships between two variables that truly exist at the individual level (e.g., *Life Satisfaction* and *Individualism*) have also been analyzed with disaggregation analyses. In such analyses, persons are the units of analysis, and country-level differences in scores are partialled out through the use of $n - 1$ dummy-coded independent variables where n represents the number of countries, and relationships between *Life Satisfaction* and *Individualism* are examined. Such analyses are sometimes referred to as least squares dummy variable (LSDV) analyses. One of the major shortcomings of such analyses is that they assume that the relationship between variables at one level of analysis (e.g., between *Life Satisfaction* and *Individualism*) is identical across units at the other level of analysis (e.g., countries). The similarity of these relationships is something that needs to be tested, not assumed.

Advocates of the LSDV approach claim that differences in relationships can be examined by including interaction terms between predictors and the dummy variables; however, such analyses pose both practical and technical (statistical) problems. Practically speaking, the analyses are cumbersome. For example, a study across 25 countries with two predictors would require regression analyses with 72 independent variables, and determining which countries have similar relationships can be difficult, if not impossible. Technically, and more important, such analyses do not conceptualize error properly. In a study such as the hypothetical example I have been discussing, there are two sources of error – the error associated with sampling individuals in each country and the error associated with sampling countries. A LSDV analysis estimates only one error term, and this error term represents an inappropriate combination of the errors from both levels of analysis. The ability to estimate two error terms simultaneously is an important advantage of MRCM over comparable OLS techniques. Moreover, as explained in the section on fixed and random effects, MRCM analyses separate true and random variability, leading to more accurate significance tests.

Despite the shortcomings of these analyses, there are OLS analyses that can take into account the multilevel nature of a cross-cultural data set in which persons are nested within countries. For example, a researcher could calculate the correlation between two variables within each of the countries in a study and then use this correlation as a dependent measure in an analysis such as an analysis of variance (ANOVA) or regression at the country level. More formally, there is a technique known as *regression by groups* in which a separate regression equation is calculated for each group (i.e., country) in an analysis. The similarity of these equations is then compared with a *F* ratio, much like the *F* ratio that tests the similarity of means in an ANOVA. Nevertheless, these analyses are not as good as a MRCM (i.e., they do not produce estimates of relationships that are as accurate based on Monte Carlo studies). This is because they do not take into account simultaneously the error associated with sampling people within each country and the error associated with sampling countries from the population of countries.

Most researchers are familiar with the error associated with selecting a sample of people. In most studies, aside from questions about the extent to which a sample is similar in important ways (e.g., sex, age, ethnicity) to the population, little importance is placed on the specific people from whom data have been collected. Samples are assumed to be random and representative of the populations from which they were sampled. Moreover, as most researchers know, statistics based on different samples drawn from the same population will be similar, but not identical. For example, the standard error of the mean describes the variance of sample means.

Most researchers are probably not as familiar with issues that arise when a sample of countries is selected and, within each of these countries, the samples of individuals are selected. In most studies, little importance is placed on the specific countries in which data have been collected (aside, perhaps, from questions of representativeness of certain cultures such as the Third World). The assumption is that countries have been sampled from the universe of possible countries. Parameters estimated from one set of countries should be similar to parameters based on another set of countries, although it is not likely that the two sets of coefficients will be identical. That is, there is some error associated with the sampling of countries. For technical reasons that are beyond the scope of this chapter, accounting for both types of sampling errors (the error associated with sampling countries and the error associated with sampling individuals) cannot be done with OLS analyses (i.e., variations of ANOVA and

regression); however, this can be done with maximum likelihood procedures, which are the basis of MRCM.

MULTILEVEL RANDOM COEFFICIENT MODELING

The principles underlying MRCM are discussed in this section in terms of the types of two-level data sets that have been discussed so far: persons nested within countries and cultures. It is possible to have more than two levels, but for an introduction and illustrative purposes, two levels will suffice. The present description of MRCM is organized in terms of the three types of research questions in which many cross-cultural psychologists are likely to be interested.

1. How can cultural-level variables explain cross-cultural differences in means of individual-level variables?
2. Controlling for cultural-level differences in means of individual-level variables, what are the within-culture (or individual-level) relationships between individual-level variables?
3. Controlling for cultural-level differences in means, how can cultural level variables explain differences across cultures in relationships between individual-level variables?

One way to think of MRCM analyses is to consider them as a series of hierarchically nested regression equations in which the coefficients from one level of analysis become the dependent measures at the next level of analysis. In essence, a regression equation is estimated for each unit of analysis (country), and the coefficients from these equations become the dependent variables in regression equations at the next level of analysis. Note that although "two-stage" OLS regression analyses may be conceptually similar in some ways to MRCM, two-stage OLS regression is not the same as MRCM, and the differences in the relationships estimated by the two techniques can be meaningful.

Such differences in estimates of relationships reflect differences in the ways in which parameters are estimated, and the more "irregular" the data are, the more accurate MRCM analyses are compared with OLS analyses. Irregular in this instance refers to the similarity across units of analysis (countries in our case) in terms of the number of observations (i.e., individuals), similarity across countries in the variances of the measures, and most important, similarity across countries of the covariances among measures (i.e., the similarity of individual-level relationships). The greater accuracy

of MRCM over OLS analyses has been demonstrated in numerous Monte Carlo studies in which random samples have been taken from populations with known parameters. In such cases, the parameter estimates provided by the MRCM techniques discussed in this chapter are meaningfully closer to the population parameters than the estimates provided by comparable OLS analyses.

Various aspects of conducting MRCM analyses are illustrated through the analyses of a hypothetical data set, presented in the Appendix. Conceptualized as a cross-cultural study, there are 10 countries (Level 2 units) and between 8 and 13 persons in each country (Level 1 units). For each country, *GDP per capita* was measured (values ranging from 8 to 16), and for each person, three measures were taken, *Life Satisfaction* (10–20), *Individualism* (1–10), and *Union Membership*, which was represented in different ways that are discussed later. There are additional variables that are discussed in other sections.

The analytic techniques described in this article are all available in the program HLM (Version 6; Raudenbush, Byrk, Cheong, and Congdon, 2004), and the analyses described in this chapter were conducted using this program. These analyses could have been conducted using other multilevel programs such as MLwiN (Rabash et al., 2000), a multilevel module in LISREL, SAS PROC MIXED (e.g., Singer, 1998), and others. Some of the terms and symbols may vary from program to program, but the terms used here should provide readers with a good introduction. Finally, many of the analytic conventions used by HLM are also used by other programs. That is, when the same models are specified, different programs should give identical results. This article describes results from HLM analyses because HLM is a popular multilevel program. HLM 6 produces two sets of results, one for robust estimates and one for nonrobust estimates. The discussion relies on nonrobust estimates because robust estimates require more Level 2 units than contained in the test data set. In HLM, the results of each analysis indicate whether robust estimates are appropriate.

It is important to keep in mind that in MRCM analyses, two parameters are (or can be) estimated for each coefficient. The first, referred to as a fixed effect, is an estimate of the central tendency (mean) of a coefficient. The questions posed by most cross-cultural psychologists concern tests of fixed effects. For example, on average, is a coefficient, such as the relationship between *Life Satisfaction* and *Individualism*, significantly different from 0? (Note that this does not examine the hypothesis that all coefficients are different from 0 or that all are less or greater than 0.) The second estimated

parameter is the random error term associated with a coefficient, and it is also tested. Is the random error for a coefficient significantly different from 0? It is common in the modeling literature to discuss coefficients as random or fixed on the basis of whether the random error term is significant (a random effect) or not significant (a fixed effect). Random error terms are discussed subsequently.

In this chapter, MRCM models and analyses are described using the nomenclature that is fairly standard for multilevel analysis. This includes specific terms (e.g., Level 1 not "lower level") and specific letters (e.g., β not b or B). Although potentially cumbersome at first, the use of these conventions facilitates communication. Multilevel analyses are inherently more complex than many single-level analyses, and the use of different terms and symbols by different authors to refer to the same entities is likely to increase readers' confusion. In standard MRCM nomenclature, for two level models, Level 1 coefficients are represented with βs (subscripted 0 for the intercept, 1 for the first predictor, 2 for the second, etc.), and Level 2 coefficients are represented with γs. As discussed later, there is a separate Level 2 equation for each Level 1 coefficient.

For pedagogical purposes, in this chapter I follow the lead of Bryk and Raudenbush (1992) and present the equations for each level separately. In fact, in MRCM analyses, all parameters in all equations (including error terms) are estimated simultaneously. Moreover, it is this simultaneity that is a part of what distinguishes MRCM from comparable OLS techniques such as two-stage least squares. OLS analyses cannot estimate more than one error term at a time for a single equation.

The first step in any MRCM analysis should be running what is called a null or totally unconditional model. These terms are used because there are no predictors at either level of analysis. Such a model is as follows:

$$\begin{aligned} &\text{Individual, Level 1:} \quad y_{ij} = \beta_{0j} + r_{ij} \\ &\text{Country, Level 2:} \quad \beta_{0j} = \gamma_{00} + u_{0j} \end{aligned}$$

In the Level 1 (individual-level) model, there are i persons for j countries on a variable y. These observations are modeled as a function of: the intercept for each country (β_{0j}, the mean of y) and deviations of each person in a country from the country mean (r_{ij}). The variance of r_{ij} is the Level 1 (or person-level or within-country) variance. There are no constraints on the similarity of the sample sizes across the countries. In the Level 2 (country-level) model, the mean of y for each of j countries (β_{0j}) is modeled as a function of the grand mean (γ_{00} – the mean of means) and deviations

of each country from the grand mean (u_{0j}). The variance of u_{0j} is the Level 2 (or country-level or between-country) variance. Such models are referred to as unconditional because y is not modeled as a function of another variable at Level 1 or at Level 2.

Although unconditional models typically do not test hypotheses, they can provide useful information. For example, they describe how much of the total variance of y is at each level of analysis. In a two-level model, the total variance is the sum of the variances of r_{ij} and of u_{0j}, and the distribution of the total variance of y suggests the levels at which further analyses might be productive. For example, if all the variance for a measure is at the person level (Level 1), it may be difficult to examine country-level differences (Level 2) in these means. The variance estimates provided by unconditional models also provide baselines that can be used to estimate effect sizes, which are discussed subsequently. The unconditional model of *Life Satisfaction* produced the following results (rounded to two decimals): $\gamma_{00} = 15.04$, Level 1 variance = 3.05, Level 2 variance = 0.65.

Analyzing Country-Level Differences in Means

A simple extension of the basic unconditional model is adding a predictor at Level 2. For example, a researcher might be interested in the relationship between mean *Life Satisfaction* (y) in a country and a country's per capita gross domestic product (*GDP*). Basically, are people living in more prosperous countries more satisfied with their lives? The equations for such a model are as follows:

$$\begin{aligned} \text{Level 1:} \quad & y_{ij} = \beta_{0j} + r_{ij} \\ \text{Level 2:} \quad & \beta_{0j} = \gamma_{00} + \gamma_{01}\,(GDP) + u_{0j} \end{aligned}$$

In this model, a country-level mean for *Life Satisfaction* is estimated for each country (β_{0j}), and the relationship between these means and each country's *GDP* is represented by the γ_{01} coefficient in the Level 2 equation. If the γ_{01} coefficient is significantly different from 0, then the relationship between *Life Satisfaction* and *GDP* is statistically significant. The analyses of the sample data set produced the following estimates: $\gamma_{01} = 0.027$, $t = .30$, *ns*. These results suggest that the average *Life Satisfaction* in a country is not related to a country's *GDP*. In the test data set, approximately 15% (.65/3.70) of the total variance of *Life Satisfaction* was between countries, which is probably a greater percent at the country level than is the case in many actual cross-cultural studies. In this regard, keep in mind that although small amounts

of variance at a level of analysis suggest that relationships may not exist there – they do not preclude entirely the possibility that such relationships exist.

Note that all coefficients in MRCM analyses are unstandardized. That is, the γ_{01} coefficient for *GDP* represents the expected change in *Life Satisfaction* for 1 unit change in *GDP*. The exact meaning of the γ_{00} and γ_{01} coefficients also depends on how *GDP* is centered. In this analysis, *GDP* was grand-mean centered. Interpreting coefficients, including centering options and standardization, is discussed later in the chapter.[3]

Although conceptually similar, the multilevel analysis that was just described is different from using an OLS analysis such as regression with countries as the unit of analysis to examine the relationship between *GDP* and aggregated *Life Satisfaction*. The essential difference is that the MRCM analyses take into account differences across countries in the reliability of the intercepts (means) for *Life Satisfaction*. Such differences will primarily be a function of the number of observations (people) in each country and the consistency of their responses. In MRCM, variance estimates reflect what is called *Bayes shrinkage* – a process in which unreliable coefficients (e.g., those based on a small number of inconsistent responses) are "shrunken" towards the mean coefficient. See Raudenbush and Bryk (2002) for more detail about this topic.

ESTIMATING WITHIN-COUNTRY RELATIONSHIPS

Predictors can also be added to the Level 1, within-country model. For example, within-country relationships between *Life Satisfaction* and *Individualism* could be examined with the following model:

$$\begin{aligned} \text{Level 1:} \quad & y_{ij} = \beta_{0j} + \beta_{1j}\,(\textit{Individualism}) + r_{ij} \\ \text{Level 2:} \quad & \beta_{0j} = \gamma_{00} + u_{0j} \\ & \beta_{1j} = \gamma_{10} + u_{1j} \end{aligned}$$

In the Level 1 (within-country or person-level) model, β_{1j} is a coefficient, called a slope to distinguish it from an intercept, representing the relationship between *Individualism* and *Life Satisfaction*, and a slope is estimated for each country. The hypothesis that the mean relationship (the mean slope) is different from 0 is represented at Level 2 by the γ_{10} coefficient. If this

[3] Technically, any type of measure can be included at Level 2, country-level aggregates, categorical variables, and so forth. Researchers will need to make informed decisions about measures that make sense given their questions of interest.

coefficient is significantly different from 0, then the mean slope is different from 0.

The test data set contains a measure, *Individualism* (*Individ*), and *Individualism* was entered group-mean centered (centering is described subsequently). The analyses of the test data set produced the following results: for the intercept, $\gamma_{00} = 15.02$, Level 1 variance = .30, Level 2 variance = 0.94, and for the slope, $\gamma_{10} = 0.85$, $t = 5.13$, $p < .001$. On average, *Life Satisfaction* and *Individualism* were significantly related. Similar to the previous example, these coefficients are unstandardized. This means that the γ_{10} coefficient represents the change in *Life Satisfaction* associated with a 1.0 unit increase in *Individualism*. In the present example, this means that for every 1.0 *Individualism* increased, *Life Satisfaction* increased .85, *on average*. Note that the Level 1 variance is substantially lower than in this analysis than it was in the unconditional model (.30 vs. 3.05). The differences in these variance estimates can be used to estimate effect sizes (and by extension, average within-country correlations), a topic discussed later. Finally, a more detailed interpretation of this slope and the intercept depends on how *Individualism* is centered.

EXAMINING DIFFERENCES BETWEEN COUNTRIES IN WITHIN-COUNTRY RELATIONSHIPS

In the analysis presented in the previous section, the hypothesis being tested concerns the average, or mean, relationship between *Life Satisfaction* and *Individualism*. It is entirely possible that the relationship between *Life Satisfaction* and *Individualism* may vary across countries. Examining differences between countries in the relationship between *Life Satisfaction* and *Individualism* could be examined with the following models. As in the first example, the country-level variable of interest is *GDP*.

$$\begin{aligned} \text{Level 1:} \quad & y_{ij} = \beta_{0j} + \beta_{1j}\,(\textit{Individualism}) + r_{ij} \\ \text{Level 2:} \quad & \beta_{0j} = \gamma_{00} + \gamma_{11}\,(\textit{GDP}) + u_{0j} \\ & \beta_{1j} = \gamma_{10} + \gamma_{21}\,(\textit{GDP}) + u_{1j} \end{aligned}$$

Note that *GDP* is in both equations. There is broad agreement among multilevel modelers that the same predictors should be included, at least initially, in all Level 2 equations. The primary reason for this is that MRCM analyses rely on covariance matrices. If a variable is not included in one equation, the tacit assumption is that it is not significant and that there is not any meaningful covariation between the coefficients across the equations.

Nevertheless, analysts will need to make their own decisions about the coefficients retained in final models.

The analyses of the test data set produced the following results: $\gamma_{21} = -0.15$, $t = -2.83$, $p < .05$. The significant γ_{21} coefficient indicated that the relationship between *Life Satisfaction* and *Individualism* varied as a function of *GDP*. Such a situation is sometimes referred to as a *cross-level interaction* or a moderating relationship because a relationship at one level of analysis varies, or is moderated by, a variable at another level of analysis.

Interpreting such a relationship is aided considerably by the calculation of estimated values. In the present example, the mean slope (the average coefficient between *Life Satisfaction* and *Individualism*) was .85. The coefficient representing the strength of the moderating relationship of *GDP* was −.15. Thus, for a country that was +1.0 above the mean on *GDP*, the estimated slope would be $.85 + 1 * (-.15) = .70$. In contrast, for a country that was –1.0 below the mean on *GDP*, the estimated slope would be $.85 - 1 * (-.15) = .90$. Note that for these analyses, *GDP* was grand-mean centered.

CENTERING

Centering refers to the reference value from which deviations of predictors are taken, and analysts need to choose centering methods carefully. Centering changes the meaning of coefficients and can change estimates and significance tests of both fixed and random effects. For analysts whose primary experience is OLS regression, it may be difficult at first to understand and appreciate the importance of centering in multilevel modeling. OLS regression analyses are almost invariably mean centered – the intercept represents the expected score for an observation at the mean of a predictor or set of predictors. Other options exist in MRCM. At Level 2 (the country level for our purposes), there are two options: uncentered (also called zero mean centering) and grand-mean centered. At Level 1 (the within-country level), there are three options: uncentered (or zero mean centered), group-mean centered, and grand-mean centered. Regardless of the level or the type of centering, analysts are strongly encouraged to generate predicted values to interpret the coefficients estimated in their analyses.[4]

[4] The way in which variables are centered varies considerably across software packages. For example, in HLM, the program centers variables automatically – analysts do not need to create any transformed variables. In contrast, when using SAS PROC MIXED, in some cases, analysts need to create centered variables before analysis. Analysts are encouraged to consult manuals for the software they intend to use to determine how variables are centered in the software they will be using.

Grand-mean centering at Level 2 is conceptually similar to the centering that is done in OLS regression. The intercept represents the expected value for the dependent measure (which could be an intercept or a slope) for a Level 2 unit (country in our case) that is at the mean of the predictor or predictors, just as it is in OLS regression.

When Level 2 predictors are uncentered, the intercept represents the expected value for the dependent measure when a predictor is 0. For example, assume that countries are classified in terms of the nature of their governments, with 0 = dual-party system and 1 = multiparty system. If this variable, which we will call *MParty*, is included uncentered, then the intercept will represent the expected value for countries for which *MParty* = 0, and the coefficient for *MParty* will represent the difference between the two types of countries. If the coefficient for *MParty* is significantly different from 0, then there is a difference in *Life Satisfaction* between countries with the two types of political systems. An analysis of *Life Satisfaction* with *MParty* as an uncentered Level 2 predictor is represented as follows.

$$\begin{aligned} &\text{Level 1:} \quad y_{ij} = \beta_{0j} + r_{ij} \\ &\text{Level 2:} \quad \beta_{0j} = \gamma_{00} + \gamma_{01}\,(\textit{MParty}) + u_{0j} \end{aligned}$$

This analysis of the test data set provided the following estimates: $\gamma_{00} = 14.69$, $\gamma_{01} = 0.71$, $t = 1.32$, *ns*. The γ_{00} coefficient represents the mean for countries that have a dual party system, that is, when *MParty* = 0. These results suggest that the average *Life Satisfaction* in a country does not vary as a function of a country's *MParty* system. If the coefficient for *MParty* was significant, then the difference between the estimated mean *Life Satisfaction* for dual-party countries (estimated value: $14.69 + 0 * .71 = 14.69$) and the estimated mean *Life Satisfaction* for multiparty countries (estimated value: $14.69 + 1 * .71 = 15.40$) would be significant.

Centering at Level 1 follows the same logic as centering at Level 2, but it is important to keep in mind the implications that centering at Level 1 has for what is analyzed at Level 2. This is because the coefficients that are estimated at Level 1 are "carried up" (at least conceptually) to Level 2, and exactly what is estimated at Level 1 will vary as a function of how predictors are centered. Similar to Level 2, when Level 1 predictors are uncentered, relationships between the dependent measure and deviations of the predictors from 0 are modeled. The intercept represents the expected score when a predictor is 0, and it is this score that is then analyzed at Level 2.

It makes little sense to model predictors as uncentered when 0 is not a valid value for a predictor, for example, when a predictor is measured

using a 1-10 point Likert scale that has no 0 point. In contrast, modeling predictors as uncentered may be sensible if predictors are coded variables for which 0 is a valid value (e.g., categorical variables) or for continuous variables for which 0 is a valid value. Moreover, by subtracting a constant, 0 can become a valid value for continuous variables that may not have a natural zero point. For example, if age is a predictor, an analyst can subtract a certain number of years so that a certain age is represented by 0. If such a variable is then entered uncentered, the intercept will represent the expected score for a person at that age. This is a common procedure in longitudinal studies in which a specific age is of primary importance.

Another option at Level 1 is group-mean centering. When predictors are group-mean centered, relationships between the dependent measure and deviations of the predictors from the mean of each group (Level 2 unit or country for many cross-cultural studies) are modeled. In this case, the intercept represents the expected score when a predictor is at the mean for each group. Aside from rounding error, when predictors are group-mean centered, intercepts are the same as they are when there are no predictors. In the example presented earlier, when *Life Satisfaction* was predicted by *Individualism*, which was group-mean centered, the mean intercept was 15.02, and it was 15.04 in the unconditional analysis. Group-mean centering is the option that is conceptually the closest to conducting a regression analysis for each group (for each country in the present case) and then using the coefficients from these analyses as dependent measures in another analysis, what is sometimes called two-stage regression. As noted previously, although group-mean centered MRCM and two stage-regression analyses are conceptually similar, they are not the same because of the differences between the two techniques in how parameters (including error) are estimated.

Substantively, group-mean centering Level 1 (individual-level) predictors may help alleviate concerns about cultural differences in mean responses. For example, if a researcher is concerned that there are country-level differences in acquiescence and wants to eliminate the influence of such differences on the results, group-mean centering predictors would eliminate such influences. Between country differences in mean acquiescence would not influence the estimation of intercepts or coefficients representing within-country relationships. Note that this is not the same as standardizing the measures within each country (a topic discussed below) – it is simply a way to control for Level 2 (country level) differences in the means of predictors.

The final centering option for Level 1 predictors is grand-mean centering. When predictors are grand-mean centered, relationships between

the dependent measure and deviations of the predictors from the grand mean of all observations are modeled. In this case, the intercept represents the expected score when a predictor is at the grand mean. Substantively, when predictors are grand-mean centered, the intercept for each group is functionally equivalent to an adjusted mean. In this case, adjusted refers to adjusted for between country differences in means of predictors.

There may be times when analyst wants to make such adjustments. For example, in research on what is known as the Big Fish in a Little Pond Effect (e.g., Marsh & Hau, 2003) the negative classroom level relationship between self-concept and mean classroom level ability occurs only when ability (an individual-level predictor) is entered grand-mean centered. There may be other occasions when analysts want to make such adjustments. Regardless, analysts need to be mindful of the implications of how they center predictors.

The similarity of the intercepts from analyses in which predictors are group- and grand-mean centered will depend upon how much group level means for predictors (countries in our example) vary. Using the test data set, predicting *Life Satisfaction* from *Individualism*, with *Individualism* grand-mean centered, produced the following results: for the intercept, $\gamma_{00} = 15.37$, Level 1 variance = .30, Level 2 variance = 4.20, and for the slope, $\gamma_{10} = 0.85$, $t = 5.18$, $p < .001$. Notice the large difference in the Level 2 (country-level variance) of the intercept between this analysis and the group-mean centered analysis (.94 vs. 4.20). This difference is due to the fact that Level 2 (country-level) variance in *Individualism* has been introduced into the model, and there is meaningful country-level variance in *Individualism*, 2.01, estimated from an unconditional analysis of *Individualism*. The country-level variance in *Individualism* can also be seen from the Level 2 data provided in Appendix A.

The fact that group-mean centering controls for Level 2 differences in Level 1 predictors can be illustrated by making Level 2 differences in Level 1 predictors larger. For these analyses, a new variable, *Individ2* was created. For countries a, b, c, and d, *Individ2* was the same as *Individualism* (*Individ*) from the original analyses. For countries e, f, g, h, and i, 100 was added to the original variable to create individ2. When *Life Satisfaction* was predicted by *Individ2*, and *Individ2* was group-mean centered, the results were *identical* to those from the original analyses. The country-level differences in *Individ2* did not contribute to the analyses.

In contrast, when *Life Satisfaction* was predicted by *Individ2*, and *Individ2* was grand-mean centered, the following estimates were produced: intercept of the intercept, $\gamma_{00} = 1.69$; intercept of the slope, $\gamma_{10} = .38$; Level 1 variance, 1.54; Level 2 variance of the intercept, 272.29; Level 2 variance

of the slope, .14. Given that this model estimates an intercept that is very different from the intercept from the unconditional model, an analyst would have to question the appropriateness of grand-mean centering in this case. Rarely are Level 2 differences in predictors as pronounced as the differences in this example; however, this example makes the point that such differences contribute to parameter estimates.

Nevertheless, some analysts argue that it is inappropriate to group-mean center predictors because the Level 2 variance in Level 1 predictors is not part of the model and it should be. Some suggest that when using group-mean centering, such variance can be included by including group means (i.e., country-level means in our case) as predictors at Level 2. At this point in time, it is difficult to provide a hard and fast recommendation regarding this specific issue. Some well-respected multi-level modelers (e.g., Raudenbush and Bryk) discuss group-mean centering without any mention of including group means as predictors at level 2. Moreover, there are scores of published articles that have reported analyses in which level 1 predictors have been entered group-mean centered and the group means of these predictors have not been included at level 2.

Given all this, I recommend group-mean centering continuous Level 1 predictors. Such a procedure makes the analyses similar (conceptually) to conducting individual regression equations for each country and using the within-country coefficients from such analyses in between-country analyses. Regarding the inclusion of group (i.e., country-level) means at level 2 to compensate for the fact that the country-level variance of these predictors is not modeled when level 1 predictors are group-mean centered, I will note that I do not include such means in my analyses. Nevertheless, individual analysts may read the literature and reach a different conclusion. Regardless, as the previous examples illustrate, different centering options (particularly at Level 1) can lead to very different results. Analysts may want to conduct group- and grand-mean centered analyses (and perhaps uncentered if appropriate) and compare the results, trying to understand whatever differences exist between or among the results. Nevertheless, as Bryk and Raudenbush (1992, p. 27) noted, "No single rule covers all cases," and analysts will need to make decisions about centering based on their questions of interest and the available data.

CODING AND CATEGORICAL VARIABLES AS PREDICTORS

Questions about centering naturally lead to questions about the nature of predictors. The previous discussion has concerned continuous variables

as predictors, and although many analysts may be interested in continuous measures, categorical measures such as gender at the person-level and country characteristics at the country level may also be of interest. In this section, I describe the use of categorical variables as predictors. Analyzing categorical variables as dependent measures is described in a separate, subsequent section.

First, it is important to note that multilevel programs do not distinguish categorical and continuous predictors. Predictors are predictors. Analysts need to anticipate the analyses they want to conduct, the groups they want to represent in their analyses, and the comparisons they want to make, by creating categorical variables before analyses. Moreover, the same centering options are available for categorical predictors as for continuous predictors, and the judicious combination of coding schemes and centering options provides a flexible means of estimating coefficients and testing specific hypotheses. In this section, I describe some of these combinations, although this description is not exhaustive.

To review quickly, dummy codes are variables that are coded 0 and 1, usually with 1 representing the presence of a condition. Contrast codes (and effect codes) represent contrasts, and typically, the coefficients need to sum to 1. For a dichotomous system, one category would be represented by 1, and the other by –1. If there are three categories, the first could be compared with the second two by coding the first category 2, and the second and third categories –1 and –1, and so forth. As noted in the previous section, centering changes what the intercept represents. To ease interpretation, it is often convenient to enter categorical variables uncentered.

These principles are illustrated with the test data set. *Union* is a dummy-coded variable representing whether a person is a member of a labor union. For the test data set, the initial analysis in which *Life Satisfaction* was the dependent measure and *Union* was the independent measure (entered uncentered) produced a nonsignificant random error term for *Union* ($p > .50$). Modeling random error terms is discussed later. The model was rerun with *Union* as a fixed effect, which produced the following parameter estimates: intercept, mean $\gamma_{00} = 14.68$, variance = .69; intercept of the slope, $\gamma_{10} = .70$.

Interpreting these coefficients is aided by estimating predicted values. When the variable *Union* is entered uncentered, the estimated score for nonmembers ($Union = 0$) is $14.68 + (0 * .70) = 14.68$, that is, the intercept. The estimated score for members ($Union = 1$) is $14.68 + (1 * .70) =$

15.38. The slope for *Union* is significantly different from 0 ($t = 2.41$, $p = .02$), and this means that, on average, people who are members of labor unions are more satisfied with their lives (15.38) than those who are not members (14.68). This model estimates an intercept for each country, which represents the score for nonmembers, and a slope for each country, which represents the difference between members and nonmembers.

Another way to model the difference would be to use a contrast coded variable representing the difference, such as the variable *Ucnt* in the example data set (1 = member, −1 = nonmember). A model with *Ucnt* (uncentered, as a fixed effect – no random error term) as a Level 1 predictor produced the following parameter estimates: intercept, mean $\gamma_{00} = 15.03$, variance = .69; intercept of the slope, $\gamma_{10} = .35$. Interpreting these coefficients is aided by estimating predicted values. When the variable *Ucnt* is 1 (for members), the estimated score is $15.03 + (1 * .35) = 15.38$. When the variable *Ucnt* is −1 (for nonmembers), the estimated score is $15.30 + (-1 * .35) = 14.68$. The difference between members and nonmembers, .70, is the same as the difference found in the previous analysis. The difference between the two analyses is what the intercept and slope for union membership represent, something that matters when differences in the intercept is modeled at Level 2.

The importance of this difference can be illustrated when *Life Satisfaction* is modeled as a function of *Union* at Level 1, and both of these coefficients are then modeled as a function of *GDP* at Level 2.

$$\begin{aligned} \text{Level 1:} \quad & y_{ij} = \beta_{0j} + \beta_{1j}\,(\textit{Union}) + r_{ij} \\ \text{Level 2:} \quad & \beta_{0j} = \gamma_{00} + \gamma_{01}\,(\textit{GDP}) + u_{0j} \\ & \beta_{1j} = \gamma_{10} + \gamma_{11}\,(\textit{GDP}) + u_{1j} \end{aligned}$$

This analysis produced the following estimates: mean intercept, $\gamma_{00} = 14.56$, a significant relationship (slope) between *Union* and the intercept, $\gamma_{01} = .24$ ($t = 2.46$, $p > .05$), a significant mean slope (*Union* effect), $\gamma_{10} = .80$ ($t = 3.35$, $p > .01$), and a significant relationship between the *Union* effect (slope) and GDP, $\gamma_{11} = -.29$ ($t = 2.93$, $p > .01$). What is important about these results is that the intercept represents the *Life Satisfaction* for people who are not members of a union (i.e., when *Union* = 0). Therefore, the significant relationship between *GDP* and the intercept ($\gamma_{01} = .24$) represents a significant relationship between *GDP* and *Life Satisfaction* for nonmembers, not for all the members of a country.

If *Life Satisfaction* is modeled as a function of *Ucnt* (the contrast variable) at Level 1 and both of these coefficients are then modeled as a function of *GDP* at Level 2, slightly different results occur.

$$\begin{aligned} \text{Level 1:} \quad & y_{ij} = \beta_{0j} + \beta_{1j}\,(Ucnt) + r_{ij} \\ \text{Level 2:} \quad & \beta_{0j} = \gamma_{00} + \gamma_{01}\,(GDP) + u_{0j} \\ & \beta_{1j} = \gamma_{10} + \gamma_{11}\,(GDP) + u_{1j} \end{aligned}$$

This analysis produced the following estimates: mean intercept, $\gamma_{00} = 14.95$, a nonsignificant relationship (slope) between *Ucnt* and the intercept, $\gamma_{11} = .09$ ($t = 1.03$), a significant mean slope (*Ucnt* effect), $\gamma_{10} = .40$ ($t = 3.35$, $p > .01$), and a significant relationship between the *Ucnt* effect (slope) and GDP, $\gamma_{11} = -.145$ ($t = 2.93$, $p > .01$). What is important about these results is that the intercept represents the mean *Life Satisfaction* for people who are neither members nor nonmembers of a union (i.e., when *Unct* = 0), a sort of "average" person.

Comparing the two results reveals that analyses of the two slopes are identical when the fact that *Ucnt* is a contrast variable is taken into account – after all, the contrast variable is simply the dummy variable with 1 subtracted from the 0s for nonmembers. The two codes are correlated 1.0 and are mathematically equivalent, so the results should be the same. The difference is in the analyses of the intercepts, and the difference in the relationships between *GDP* and the intercepts in the two analyses suggests that *GDP* may be related differently to the *Life Satisfaction* for members and nonmembers.

Such a possibility can be examined directly by conducting an analysis in which *Life Satisfaction* is modeled as a function of two dummy codes, one representing union members and the other representing nonmembers. In the test data set, the dummy coded variable for nonmembers is variable *NUn*. Note that in this model, the intercept is deleted, and such models are sometimes referred to as "zero or no intercept models." In such analyses, the coefficients represent the means for Level 1 categories, and in the present example, the coefficient *Union* represents the mean for members, and *NUn* represents the mean for nonmembers.

$$\begin{aligned} \text{Level 1:} \quad & y_{ij} = \beta_{1j}\,(Union) + \beta_{1j}\,(NUn) + r_{ij} \\ \text{Level 2:} \quad & \beta_{1j} = \gamma_{10} + \gamma_{11}\,(GDP) + u_{1j} \\ & \beta_{2j} = \gamma_{20} + \gamma_{21}\,(GDP) + u_{2j} \end{aligned}$$

This analysis produced the following estimates: mean intercept for *Union*, $\gamma_{10} = 15.36$, and a nonsignificant relationship (slope) between *GDP* and the *Union* coefficient, $\gamma_{11} = -.05$ ($t < 1$), mean intercept for *NUn*,

$\gamma_{20} = 14.55$, and a significant relationship between *GDP* and this coefficient, $\gamma_{11} = .24$ ($t = 2.53$, $p < .05$). In other words, *GDP* was related to *Life Satisfaction* for nonmembers, but was not related to *Life Satisfaction* for union members.

Using such dummy codes, means (e.g., γ_{10} and γ_{20}) can be compared using tests of fixed effects as described subsequently. Although the results of such comparisons of Level 1 coefficients representing means will typically be similar to the results of significance tests of Level 1 coefficients representing differences between categories, the results may not be exactly the same. This is because when differences are modeled at Level 1 with a contrast variable or a single dummy code, the model estimates a difference score between (or among) categories for each country and then estimates mean difference scores. When dummy codes are used with a no intercept model, means for each category are estimated for each country, and then differences among these means are tested. Note that this type of dummy-coded analysis is possible only when observations can be classified using a mutually exclusive system, that is, an observation falls into one and only one category. The number of categories that can be represented is not limited technically, but each Level 1 observation must be classified as belonging to one and only one category.

There is also an important caveat regarding relationships between Level 2 variables and Level 1 slopes (which in this instance represent means). The coefficients estimating means for each category are stable even when some Level 2 units do not have observations in all categories. For example, assume a three-category system in which 50% of participants do not have observations in the third of these categories. The coefficient (mean) for Category 3 estimated using all participants will be the same as a mean estimated from an analysis that includes only those who have some observations in Category 3; however, estimates of relationships between Level 2 variables and means for this category will not be the same. When a substantial number of Level 2 units (perhaps 10% or more) are missing observations in a category, analysts should conduct separate analyses on subsets of countries that have observations in all categories and those that do not to determine whether the subsamples differ meaningfully in other ways.

Analysts may also be interested in nonexclusive, overlapping categories at Level 1, for example, union membership and gender. One way to deal with such categories is to combine them into mutually exclusive categories (e.g., male members, female members, male nonmembers, and female nonmembers) and then use dummy codes for each of the resultant categories and proceed as described earlier. This may not always be practical or

desirable, and categorical predictors can be represented with contrast (or effect) codes, as in the sample data set variable *Ucnt*.

An important advantage of contrast coding is that it allows level 2 (country-level in our case) differences in difference scores to be modeled. Moreover, multiple contrasts can be included simultaneously, including contrasts when there are more than two categories. When using multiple contrast codes, analysts need to be mindful of the fact that the coefficients are adjusted for each other, meaning that the estimate of a specific contrast may vary as a function of the other contrasts in a model. A disadvantage of contrast coding is that it does not allow for examination of differences in relationships between Level 2 variables and Level 1 category means.

Categorical codes can also be used to adjust for country-level differences in the distribution of within-country groups. If a categorical variable (dummy or contrast code) is entered "grand-mean centered," then the intercept represents the country-level mean adjusted for country-level differences in the relative frequency of the categorical variable. For example, in the test data set, when *Union* is entered grand-mean centered as a predictor of *Life Satisfaction*, the intercept for *Life Satisfaction* is 15.05, the same (within rounding) as it was from the unconditional model. By grand-mean centering *Union*, country-level differences in this variable were eliminated from the analysis.

I recommend that analysts prepare both dummy and contrast codes for categorical variables and model dependent measures in different ways. If contrast- and dummy-coded analyses provide dramatically different conclusions about mean differences, this should be investigated because they should not. The two types of coding provide different advantages, and analysts will need to understand when to use one type or the other. For example, at Level 1, contrast-coding groups allow analysts to examine between-country differences in within-country differences. Continuing the previous example, do differences between union members and nonmembers in work attitudes vary as a function of a country-level characteristic such as GDP? Dummy-coding groups allows analysts to determine whether relationships between country-level measures and within-country means vary across groups within each country. For example, do relationships between work attitudes and GDP differ between union members and nonmembers?

I think the use of the types of categorical codes I have described is one of the most underutilized and potentially powerful aspects of multilevel modeling. Through the creative and judicious use of different types of coding combined with different types of centering, analysts can estimate precise (in terms of the relationships they represent) parameters. Such

estimates can then be compared using the test of fixed parameters discussed elsewhere in this chapter. The critical step in this process is to anticipate the exact analyses that are to be done and prepare variables that represent the parameters of interest. This topic is discussed in more detail in Nezlek (2001, 2003).

FIXED AND RANDOM EFFECTS

One of the advantages of the maximum-likelihood procedures used in MRCM is the separation of true and error (or random) variance, a separation similar to what is done in structural equation modeling. In contrast, in OLS analyses, there is only one variance estimate – true and random variance are not separated. In the results of a MRCM analysis, this separation is indicated by the fact that, as mentioned earlier, for each variable in a Level 1 equation, two terms are estimated, usually called a fixed effect and a random effect. The fixed effect, which is used to test whether a coefficient is significantly different from 0, has a variance (usually labeled a standard error), and there is a separate estimate of the random variance, usually called a random effect. The separation of true and random variance improves the accuracy of the significance tests of effects, part of what makes MRCM a better way to analyze multilevel data sets than comparable OLS analyses.

For most purposes, researchers will be interested in the fixed effects of the predictor variables included in their models. For example, the significance test of the slope between *Life Satisfaction* and *Individualism* (was it different from 0?) in the previous example was a test of the fixed effect. Like fixed effects, random effects are tested for significance. Is a random effect significantly different from 0? Unfortunately, some researchers interpret the significance of a random error term as a test of whether units of analysis vary. They assume that if the random variance associated with a slope is not significant, then all units of analysis have the same slope. For the example data set, this would mean that if the random variance associated with the slope between *Life Satisfaction* and *Individualism* was not significant, then it could be assumed that all countries had the same slope, the same relationship between *Life Satisfaction* and *Individualism*.

Although intuitively appealing, this is not quite true. Technically speaking, the significance test of the random effect associated with a coefficient indicates if true and random variance can be reliably separated. It does not formally test whether Level 2 units vary in some way, for example, do all countries have the same slope? Admittedly, the lack of a significant random error term suggests that there may not be a lot of variance in a coefficient.

The absence or presence of a random error term needs to be understood within the context of random and nonrandom variation. Coefficients that have a random error term are described as "randomly varying" or as random coefficients. Coefficients that have do not have a random error term can be what is described as *nonrandomly varying*. In the previous example, the *Satisfaction–Individualism* slope was modeled as randomly varying, that is, there was a significant random error term estimated for the slope. If the random error term is deleted, the slope is "fixed" (and is called a fixed coefficient), although one can still model variability in the slope at Level 2, and such a slope would be termed *nonrandomly varying*.

Repeating the analyses of the *Satisfaction–Individualism* slope with the slope fixed (i.e., the random error term was deleted) and *Individualism* entered group-mean centered, produces the following results. The intercept of the slope (γ_{10}) with no Level 2 predictors was .68, different from the slope when the random error term was included (.85 from the previous analyses). It is not unusual for the fixed part of a coefficient to change when a random term is eliminated. Moreover and more important, the nature of this change cannot be predicted. Slopes that were significant when modeled as random can be nonsignificant when modeled as fixed, and vice versa.

When *GDP* was included at Level 2, the γ_{11} coefficient representing the moderating relationship was $-.08$, $p = .09$. That is, fixing the slope did not prevent modeling variability in the slope. The ability to model the variability in this slope without modeling it as random coefficient was not a function of the fact that there was a significant random error term when the slope was modeled as random. Variability in slopes can be modeled even when they do not have a significant random error term.

The meaning of fixing a coefficient can also be understood by looking at estimated values for coefficients. In HLM, these are in residual files, and for the test data set, the Level 2 residual file contains the country-level estimates of the intercepts and slopes (and other statistics that are not relevant at this point). For the analysis in which slope for *Individualism* was fixed and *GDP* was not included as a predictor, the "fitted value" for the slope was .68 for all the Level 2 units (i.e., all countries). Some argue that this means that all countries had the same slope; however, this is not exactly true. When *GDP* was included as a predictor, the fitted values for the *Individualism* slope varied across countries. They varied nonrandomly as a function of *GDP*. When the slope was fixed and *GDP* was not included, the variability among the slopes was not being modeled. Not modeling the variability is not the same as saying that the slopes did not vary in some way.

It is inappropriate to conclude on the basis of a nonsignificant random error term that a Level 1 coefficient (usually a slope) does not (or can not) vary. To prove that a slope does not vary at all would require modeling the coefficient with an infinite number of Level 2 predictors. Short of this, a nonsignificant random error terms means that a coefficient does not vary randomly. A significant random error term means that a coefficient varies randomly, which formally means that there is enough information to separate true and random variability for that coefficient. For researchers interested in variability per se, the presence of a significant random error term means that Level 2 units (usually countries for a cross-cultural researcher) vary; however, the absence of a significant random error term does not mean that they do not vary.

Although random error terms typically do not test hypotheses per se (at least for many cross-cultural researchers), they must be properly specified before examining significance tests of fixed effects. The *error structure* (as the covariance matrix of random terms is called) must be specified properly because an improper error structure creates a "misspecified" model, which in turn, can lead to inaccurate significance tests of the fixed effects. Moreover, the direction of this inaccuracy cannot be predicted. That is, fixing an effect that should be modeled as random (deleting a random term that should be included) can make the fixed part of a coefficient significant when it should not be, or vice versa, just as including a random effect that should not included. Practically speaking, the manner in which random error terms are specified varies dramatically across software packages, so much so that describing the various options is well beyond the scope of this chapter. It is worth noting that in HLM, by default, predictors are entered as fixed: Analysts need to "make them random" explicitly.

Conceptually, most coefficients in cross-cultural studies should probably be modeled as random – the countries have been randomly sampled from a population of countries, and this sampling needs to be represented. Nevertheless, the data may not be able to estimate reliably all the random error terms in a model and the covariances among these random error terms. Most multilevel modelers argue that nonreliable error terms should be eliminated, although a minority argue that some estimate of the random error should be made based on information from other sources, for example, previous studies. At this point, most researchers will be on solid ground if they eliminate unreliable random error terms from their models, keeping in mind that fixed coefficients can vary nonrandomly.

Also, the norm among multilevel modelers is to use a more generous probability level than .05 when making decisions about random error terms.

In my own work, I allow error terms that are significant at .10 to remain in a model and delete those that are above .15. When *p* values are between .10 and .15, I run models with and without the error term to see the impact of including or excluding the error term. Decisions about random error terms in this "gray area" can also be informed by comparing error covariance matrices from different models using goodness of fit indices.

A discussion of guidelines for making decisions about random error terms is provided in Nezlek (2001). In that article, I discussed three bases for making decisions about modeling coefficients as fixed or random: theoretical, statistical, and practical. Theoretically (or conceptually), it is possible (although not typical) that some coefficients should be fixed because they have a narrow *breadth of inference* or *inference space*, that is, they are meant to describe a very specific population. As already discussed, coefficients can be fixed if the random error term is not significant (statistical). Finally, coefficients may be fixed if estimating them (and their covariances with other error terms) prevents a model from converging, a practical issue. In this regard, many multilevel modelers look for models to converge in less than 500 iterations.

To me, modeling and interpreting error structures within the multilevel framework is perhaps the most puzzling aspect of planning and interpreting task multilevel analyses. What does a nonsignificant random error term mean? What do correlations between random error terms represent? Articles and books can provide seemingly (or actually) conflicting advice and interpretations regarding such topics. Even within the community of scholars who study such techniques per se, there is far from a consensus regarding how to interpret error within the multilevel framework. In this chapter, I have provided what I think is sound advice regarding the interpretation of error variances; this advice is based on my reading of the literature and, more important, on my experience analyzing multilevel data structures of all sorts. Other scholars may provide different advice, and at this point, analysts may need to consult various sources and make decisions for themselves.

INTERACTIONS

Within MRCM, understanding statistical interactions is a bit more complex than it is within single level analyses because interactions can be either within or between levels or can blend the two. One of the simpler forms is the between-level interaction, often referred to as a cross-level interaction,

which represents a type of moderated relationship. A cross-level interaction occurs when a Level 1 relationship (a slope) varies as a function of a Level 2 variable. The example from the test data set showing how the relationship between *Life Satisfaction* and *Individualism* varied as a function of *GDP* represents a cross-level interaction.

Within-level interactions at Level 2 are fairly straightforward. Setting up and interpreting within-level interactions at Level 2 (the country level for our purposes) is similar to setting up and conducting interactions in OLS regression (see Aiken & West, 1991). Continuous measures are mean-centered and then cross-multiplied with other continuous measures or categorical measures and interactions are interpreted by estimating predicted values, typically for observations $\pm$ 1 *SD* from the mean for continuous measures and for observations in each group for categorical measures.

Within-level interactions at Level 1 are somewhat more complex, although the logic is the same. Most important, before creating the products representing the interaction terms, continuous variables should be centered within each group (i.e., within each country). For example, in the sample data set, to create an interaction involving *Individ*, the country mean for *Individ* would need to be subtracted from the raw *Individ* score within each country, and this centered score would be multiplied by the other variable involved. For Country A, this would be 2.75, for Country B, it would be 3.67, and so forth. The resulting interaction terms should be entered *uncentered* into the model because the terms used to create them were centered when they were created. Other terms would be entered group-mean centered. This makes generating predicted values easier.

Analysts who are interested in generating predicted scores $\pm$ 1 *SD* (the standard) need to exercise care when using predicted scores to understand Level 1 interactions. When doing so, it must be kept in mind that within-country *SD*s must be used, and these must be generated using variance estimates from unconditional models. Moreover, Level 1 interactions may also vary as a function of Level 2 variables. In such cases, different sets of Level 1 coefficients need to be generated representing countries at Level 2. This could entail different groups of countries at Level 2 or countries that are $\pm$ 1 *SD* on a Level 2 variable. Analyses of interactions within Level 1 and modeling of Level 2 differences in such interactions can be found in Nezlek and Plesko (2003). Resources for evaluating interactions within the multilevel context can also be found in Preacher, Curran, and Bauer (2006).

MODERATION AND MEDIATION

Similar to understanding interactions, understanding moderation and mediation within the multilevel framework is conceptually similar to understanding moderation and mediation within the single-level framework. Between levels, moderation can take the form of a cross-level interaction and can be examined through significance tests of Level 2 predictors of Level 1 slopes. In fact, cross-level interactions are sometimes referred to as moderating relationships because a Level 1 relationship varies as a function of, or is moderated by, a Level 2 variable. This was illustrated in the previous example in which cultural differences in individual-level relationships between *Life Satisfaction* and *Individualism* were modeled as a function of *GDP*.

Within levels, moderation can be examined by examining representing interactions of predictors much like what is done in OLS regression. For example, an interaction term between the country-level variables *MParty* and *GDP* could be created to determine if relationships between mean *Life Satisfaction* and *GDP* varied as a function of the political system in a country. Within Level 2, interpreting the results of such analyses is much like interpreting the results of OLS analyses: Significant interaction terms indicate moderation.

Evaluating moderation within Level 1 is similar, but somewhat more complex. Similar to Level 2, you need to create interaction terms representing the combined effects of the two variables in question. For example, union membership and individualism could be combined to determine if the relationship between *Life Satisfaction* and *Individualism* was similar for members and nonmembers. In this example, it would probably be best first to center *Individualism* around the mean for each country then multiply these centered values by the union membership variable, and then enter the membership variable and interaction terms uncentered. If the interaction term is significant, one can conclude that the relationship between *Life Satisfaction* and *Individualism* for union members is different than it is of nonmembers. The exact nature of this difference could be determined by generating predicted values representing relationships for members and for nonmembers.

The tricky part of such an analysis is the possibility that coefficients representing moderating effects will vary across Level 2 units. The significance tests of the fixed effect of interaction terms representing a moderation effect test if the mean Level 1 moderation effect is significantly different from

0. It is possible that moderation may be stronger or weaker for different Level 2 units (countries for our purposes). For example, the difference between union members and nonmembers in relationships between *Life Satisfaction* and *Individualism* may be larger in some countries than it is in others. Moreover, the likelihood of such a possibility is not indicated by the significance test of the random error term of the Level 1 moderating effect. For example, Nezlek and Plesko (2003) found that Level 2 variables moderated Level 1 moderating effects even when the coefficient representing a Level 1 moderating effect had no significant random error term.

Evaluating mediation within multilevel models is not that well understood, and thus the following discussion should be treated as somewhat tentative. I think a good starting point is the traditional OLS rule as discussed by Baron and Kenny (1986). Mediation occurs when the relationship between an outcome and a predictor is rendered insignificant by including a second predictor that is itself related to the outcome and to the first predictor. Informed (albeit technically focused) discussions of mediation and suggestions for ways to evaluate mediation within the multilevel context can be found in Bauer, Preacher, and Gill (2006) and Krull and MacKinnon (2001).

At Level 2, it seems that the OLS rule discussed by Baron and Kenny can be applied relatively straightforwardly. In terms of the example I have been discussing, the relationship between *GDP* and *Life Satisfaction* might be mediated by another country-level measure such as the ease with which people in a society can change social statues (*Ease*). Assume that *Ease* is also related to *Life Satisfaction*, and at the country level, *Ease* is related to *GDP*, and when *Ease* is included in the Level 2 equation predicting *Life Satisfaction*, *Ease* is significant and *GDP* is not. Such a result suggests that *Ease* mediates the relationship between *Life Satisfaction* and *GDP*.

Within Level 1, the situation is much more complex, in part because Level 1 mediation may vary across Level 2 units of analysis. For example, let us assume that we measure an additional individual-level variable, *Self-esteem*, and we find that *Self-esteem* is related to both *Individualism* and *Life Satisfaction*. Furthermore, when *Life Satisfaction* is predicted by both *Self-esteem* and *Individualism*, the coefficient for *Self-esteem* is significant, whereas the coefficient for *Individualism* is not. This is a prima facie case for mediation.

Nonetheless, it is entirely possible that such mediation does not occur in all countries. As discussed previously, tests of fixed effects of Level 1

coefficients concern the average or mean coefficient. This leaves open the possibility that countries may have a pattern of relationships that differs from the mean, a situation that is sometimes referred to as *moderated mediation*. A small caveat is in order at this time. Some explanations of Level 1 mediation (e.g., Kenny, Korchmaros, & Bolger, 2003) rely on significance tests of the random error terms associated with slopes to determine if coefficients representing mediational relationships vary across Level 2 units. Such explanations do not consider the possibility that slopes representing medational effects may vary nonrandomly. Such problems are particularly salient for cross-cultural researchers who often have a limited number of Level 2 units (countries) available. Small numbers of Level 2 units makes it difficult to estimate random error terms. If researchers rely solely on the presence of significance of random error terms to make judgments about whether Level 1 coefficients vary across countries, they may be drawing inaccurate conclusions. Moreover, as noted later, there is some confusion regarding the meaning of changes in Level 1 residual variances as a means of evaluating effect sizes. To the extent that explanations of mediation rely on changes in Level 1 variances, such explanations need to be evaluated cautiously.

At present, it is difficult to provide unambiguous guidelines regarding lower level mediation in multilevel models. The traditional OLS rule (e.g., Baron & Kenny, 1986) seems like a good place to start because it relies on significance tests of fixed effects rather than on variance estimates. Regardless, analysts need to be aware of the possibility that all of the relationships needed to establish mediation may vary across Level 2 units. That is mediation may exist in some countries but not in others.

COMPARING COEFFICIENTS: TESTS OF FIXED EFFECTS

Although the primary emphasis in much published research is on significance tests of individual coefficients, in MRCM coefficients can be compared. Such comparisons can involve slopes or intercepts, and depending on the sophistication of the analyst, these tests can be powerful. These tests (called tests of fixed effects in HLM) rely on the impact of constraints on a model. The procedure is the same as testing constraints within structural equation modeling. An analyst specifies a constraint, for example, the difference between two slopes is 0, and the impact of the fit of this constraint on a model is evaluated. If the constraint leads to a poorer fit, the hypothesis underlying the constraint is rejected, for example, the difference between the slopes is not 0 – the slopes are different.

For example, as represented in the following model, assume we are interested in the individual-level relationships between *Life Satisfaction* and *Individualism* and *Self-esteem*.

$$\begin{aligned}
\text{Level 1:} \quad & y_{ij} = \beta_{0j} + \beta_{1j}\,(\textit{Individualism}) + \beta_{2j}\,(\textit{Self-esteem}) + r_{ij} \\
\text{Level 2:} \quad & \beta_{0j} = \gamma_{00} + u_{0j} \\
& \beta_{1j} = \gamma_{10} + u_{1j} \\
& \beta_{2j} = \gamma_{20} + u_{2j}
\end{aligned}$$

The relative strength of the relationships between *Life Satisfaction* and *Individualism* and *Self-esteem* can be examined by comparing the γ_{10} and γ_{20} coefficients, representing the mean slope for *Individualism* and *Self-esteem* respectively. This is done by imposing a constraint on the model – in this instance, constraining the difference between γ_{20} and γ_{30} to be 0. If the constraint leads to a significant decrease in the fit of the model, then one can conclude that the coefficients (the mean slopes) are not the same. The significance test is a chi-squared with 1 degree of freedom. Note that such a procedure can be used to compare the means for *Union* and *Non-Union* in the examples using dummy coded predictors.

These procedures are flexible and can be used to compare various types of relationships. For example, assume that a Level 2 predictor of the slopes is added to the previous model. *GDP* is a Level 2 predictor that is being used to model country-level differences in the *Individualism* and *Self-esteem* slopes.

$$\begin{aligned}
\text{Level 2:} \quad & \beta_{0j} = \gamma_{00} + \gamma_{10}\,(\textit{GDP}) + u_{0j} \\
& \beta_{1j} = \gamma_{10} + \gamma_{11}\,(\textit{GDP}) + u_{1j} \\
& \beta_{2j} = \gamma_{20} + \gamma_{21}\,(\textit{GDP}) + u_{2j}
\end{aligned}$$

The strength of the moderating relationship of *GDP* on *Individualism* and *Self-esteem* can be compared by comparing the γ_{10} and γ_{20} coefficients. These two coefficients represent the moderating relationships of *GDP* for *Individualism* and *Self-esteem*, respectively.

These examples have focused on comparisons of only two coefficients, but constraints can involve more than two coefficients. For example, assume there are three Level 1 predictors in a model. This would generate four fixed effects at Level 2, one for the intercept (γ_{00}) and one for each of the three slopes (γ_{10}, γ_{20}, and γ_{30}). A constraint could compare the average of the first two slopes with the third (γ_{10}, γ_{20}, vs. γ_{30}), which could be coded -1, -1, 2. Similarly, the moderating relationship of Level 2 variable on each of these slopes could be compared.

Constraints can also have more than 1 degree of freedom. Assume the broad hypothesis is that three slopes, X1, X2, and X3, are different, much like the null hypothesis of an ANOVA with more than two groups. This hypothesis could be tested with a 2 degree of freedom constraint, perhaps coded as 1, −1, 0, and 1, 0, −1 for γ_{10}, γ_{20}, and γ_{30} respectively. Keep in mind that just like the *F* test from an ANOVA, if such a constraint were significant, an analyst would not know exactly which slopes were different from each other.

These examples represent only a small sample of the possible applications of this technique. I believe tests of fixed effects are underutilized in research using MRCM. With appropriate coding schemes and constraints, analysts can use MRCM to conduct ANOVA-like analyses for categorical variables, they can compare the strength of relationships without having to rely on variance estimates of questionable meaning, and so forth. One reason I have emphasized the careful understanding of exactly what each coefficient in a model represents is to encourage analysts to create models in anticipation of using these tests.

STANDARDIZATION OF MEASURES

Contrary to the experience of analysts who have worked primarily with OLS regression, MRCM estimates only *unstandardized* coefficients. MRCM analyses estimate unstandardized coefficients because the algorithms rely on covariance, rather than correlation, matrices to estimate parameters. Although this is desirable from a modeling perspective (covariance matrices have more information than correlation matrices), it is not necessarily desirable from a substantive perspective. That is, frequently, researchers are interested in examining relationships in which the influence of differences in the variances of measures are controlled or eliminated, rather than maintained.

Although standardization per se is not possible, standardized coefficients can be indirectly estimated in some cases. This is easiest at Level 2. If Level 2 measures are standardized before analysis, coefficients representing relationships between Level 2 variables and Level 1 coefficients will represent the change in a Level 1 coefficient associated with a standard unit increase in the Level 2 predictor. Technically, the analysis will still estimate unstandardized coefficients – the change in a Level 1 coefficient associated with a 1-point change in a Level 2 measure; however, because the Level 2 measure is now standardized, a 1-point change represents a standard unit.

At Level 1, the situation is not so straightforward. First, it is best to avoid such problems by designing studies so that measures have similar variances, for example, use the same scale for different measures. This is not always possible however (e.g., the data may have been collected), and so measures can be transformed to reduce or eliminate differences in variances (e.g., divided by a constant). Analysts should avoid (or consider carefully) standardizing *within* countries, that is, calculate a standard deviation for each country and represent observations in terms of standard deviations from each country's mean. Such standardization artificially sets the mean for all countries to 0, and in the process, eliminates mean differences in intercepts by eliminating any variance in the intercept.

Although changing the variance of an individual measure using a linear transformation will not change the significance tests of individual coefficients, variance differences do contribute to the significance tests of comparisons of coefficients. For example, assume a model in which there are two Level 2 predictors of a Level 1 intercept. Changing the variance of these predictors will change the results of the comparison of the Level 2 coefficients. Tests of the individual effects will not change, but chi-squared tests of constraints will. Similarly, the variances of Level 1 predictors contribute to tests of constraints of their means at Level 2.

ESTIMATING EFFECT SIZES

Within the multilevel framework, the strength of the relationships between two variables can be evaluated in two ways. The first, and least controversial, is simply to interpret the size of the coefficient. This can be illustrated using analyses of the example data set, keeping in mind that MRCM estimates unstandardized coefficients. For example, the estimated mean slope between *Life Satisfaction* and *Individualism* in the original analysis was .86. This means that *Life Satisfaction* increases .86 for every 1-point increase in *Individualism*. To estimate the increase or decrease associated with a 1 *SD* requires an estimate of the within-country *SD*. This is not done, however, using simple single-level descriptive statistics that ignore grouping.

The estimated within-country standard deviation can, and should, be derived from an unconditional analysis of the predictor. The within-country standard deviation is the Level 1 standard deviation, or the square root of the Level 1 variance if the standard deviation is not provided directly. For the sample data set, the Level 1 *SD* of *Individualism* is 2.16. By the way, simply taking the *SD* of all the Level 1 observations produces an estimate

of 2.54. Estimated *Life Satisfaction* when *Individualism* is −1 *SD* would be $15.03 - (.86 * 2.16) = 16.89$. Estimated *Life Satisfaction* when *Individualism* is –1 *SD* would be $15.03 - (.86 * 2.16) = 13.17$.

The strength of the slope can also be evaluated by estimating the percent of variance in the dependent measure accounted for by the predictor. This is done by comparing the residual Level 1 variance from the totally unconditional model of the dependent measure (3.05 for *Life Satisfaction*) to the Level 1 variance with the predictor included (.30 when *Individualism* is included). In this analysis, the two variables share just over 90% of the variance, which translates into an average within-country correlation of approximately .95. Also, reversing the process (predicting *Individualism* from *Life Satisfaction*) produces approximately the same numbers, a similarity that is not always the case.

Using residual variances to estimate the strength of relationships is a somewhat controversial procedure among multilevel modelers. For example, although Kreft and de Leeuw (1998) discussed R^2, they advised caution when interpreting such estimates of effect sizes: "In general, we suggest not setting too much store by the calculation of R^2_B [Level-2 variance] or R^2_W [Level-1 variance] (p. 119)." In part this is because adding significant Level 1 predictors does not necessarily lead to a reduction in residual variances.

Unlike OLS analyses, in which significance tests of effects are based on reductions in error variance, in MRCM analyses, significance tests of the fixed effect of a coefficient and random error terms are estimated separately. In some rare cases, adding predictors to a Level 1 model may lead to an increase in residual variance, a mathematical impossibility in OLS. There are also questions about how centering affects such estimates (Raudenbush & Bryk, 2002), with the general agreement that predictors should be group-mean centered. My advice in this regard is to follow Kreft and deLeeuw (1998) and be cautious. For the moment, it suffices to note that it appears that models with single predictors at Level 1 with no Level 2 variables provide a reasonably stable estimate of the variance shared by two Level 1 variables and that using reductions in residual variances associated with additional Level 2 predictors seems less problematic.

NONLINEAR OUTCOMES

Thus far, the discussion has assumed that dependent measures are continuous and more or less normally distributed; however, cross-cultural researchers may be interested in outcomes that are not continuous or not normally distributed, for example, categorical measures, highly skewed

count data, and so forth. Analyses of such measures rely on the same logic as analyses of measures that are normally distributed, but they use different algorithms. Different algorithms are necessary because for measures that are not normally distributed, means and variances are not independent, and this lack of independence violates a critical assumption. For example, the variance of a binomial is *npq*, where $n =$ number of observations, $p =$ the probability of the more common outcome, and $q = 1 - p$. As the mean changes (the expected value, p), the variance changes.

Analyses of nonlinear outcomes are structurally similar to the analyses of linear outcomes. A Level 1 model is estimated, and then the coefficients from Level 1 are then analyzed at Level 2. In the sample data set, there is a variable *Union*, which indicates whether a person belongs to a labor organization, and the percent of people who belong can be examined using a Bernoulli model with $n = 1$. The model is as follows:

$$\text{Prob}(\text{y} = 1|\beta_{0\text{j}}) = \phi.$$

The transformation represented in this equation normalizes Bernoulli ($N = 1$) outcomes. The null hypothesis, a mean of 0, represents 50%. The coefficient from this analysis, the log-odds of belonging, is .08 (unit-specific, nonrobust estimate), corresponds to 52%, which is this case was not significantly different from 0. Although the null hypothesis is always that a coefficient is different from 0, what 0 represents will vary as a function of the transformation being used (i.e., the type of nonlinear outcome).

When analyzing nonlinear outcomes, predictors can be added at Levels 1 and 2 just as they are added to analyses of linear outcomes, and the results of the analyses are interpreted similarly. For example, *GDP* is included in the Level 2 model (grand-mean centered), the resulting coefficient is $-.15$, which is not significant ($p = .17$). Assuming that it was, the relationship could be interpreted by generating predicted values for countries $\pm$ 1 *SD* on *GDP*. The *SD* of *GDP* is 2.63, so a country $+1$ *SD* would have a predicted log odds of $-.31 = .08 + (2.63 * (-.15))$, corresponding to 42%. A country -1 *SD* would have a predicted log odds of $.47 = .08 - (2.63 * (-.13))$, corresponding to 62%. Note that although *GDP* was grand-mean centered, the intercept in this analysis is slightly different from the intercept in the unconditional model.

When analyzing nonlinear outcomes, analysts should be aware of the following:

1. Interpreting the results of nonlinear analyses needs to be done carefully, particularly for analysts who are not familiar with logistical

regression and related techniques. Although transformations may be clearly described, producing predicted values, which are needed to understand the coefficients, can be quite complex. To generate predicted values, I use a spreadsheet with cells containing each step of the equation that is needed to generate point estimates from log odds.
2. In analyses of nonlinear outcomes Level 1 variances are not estimated. This is because of the nature of the algorithms used in these analyses.
3. For analyses of nonlinear outcomes, HLM produces two sets of coefficients, unit-specific and population-average, and such coefficients can be meaningfully different. Estimating coefficients for nonlinear outcomes requires a "link function," and different link functions are available. A detailed discussion of link functions is well beyond the scope of this chapter. Nevertheless, it may be helpful to note that unit-specific coefficients refer to relationships such as a slope (e.g., the change due to a one unit change in a predictor) at the discrete unit (e.g., country), whereas population-average coefficients refer to relationships as they exist in the population. Blanket recommendations regarding which of these is appropriate are not possible, and interested readers are encouraged to consult Raudenbush and Bryk (2002) for more detail.

INTERPRETING RESULTS

For analysts whose experience is primarily with single-level OLS analyses that produce standardized coefficients, developing a sense of how to interpret the results of MRCM analyses will take some time. Following are some important considerations.

1. MRCM analyses produce two (or more) sets of coefficients, and in the case of cross-level interactions (described subsequently), coefficients at lower levels of analysis may need to be interpreted in light of coefficients at higher levels of analysis. More levels create a more thorough understanding, but they also create more complexity.
2. Second, most coefficients in most analyses will be unstandardized. See the previous section on standardization. Although analysts can still rely on significance tests to determine whether relationships are significantly different from 0 or different from each other, the fact that coefficients are unstandardized needs to be kept in mind.
3. Significance tests of the fixed effect of a coefficient can vary as a function of the inclusion or exclusion of a random error term for that

coefficient. Before evaluating the results of significance tests of fixed effects (the tests that are most relevant for most researchers) error terms need to be specified properly. For most analyses, this will mean that error terms that are reliable should be retained, whereas those that are not reliable should be eliminated from the model, that is, the effects should be "fixed."

4. The meaning of coefficients depends on how variables are centered, and different centering options can produce different (sometimes dramatically different) significance tests. Unlike the specification of error structures, which often has a post hoc component (i.e., eliminating error terms that cannot be estimated reliably), centering is something that should be done in advance. Analysts should know in advance what they want each coefficient to represent and should select the centering options that represent these quantities.

With these considerations in mind, the key to understanding the results of MRCM analyses is to generate predicted or expected values as I have done with the analyses of the sample data set. In the case of categorical measures, this would mean expected values for each category, whereas for continuous measures, one might chose to estimate values for observations $\pm$ 1 *SD* on predictors. Given the potential complexity of the results, the importance of generating predicted values cannot be overstated. Such an emphasis contrasts sharply with the emphasis in many single-level OLS analyses on significance tests of standardized coefficients.

REPORTING RESULTS

As a guiding principle, authors need to recognize that, at least for the next few years, the majority of readers will probably not be that familiar with MRCM. Although scholarly articles are meant to educate, to expose readers to new approaches and new techniques, scholarly articles are also meant to inform, to provide readers with a better understanding of substantive issues. For most articles, this means that detailed presentations of results and aspects of the analyses that are not central to the substantive questions of the study may confuse more than they inform. Certainly, authors need to provide sufficient detail so that readers (and reviewers!) can understand what was done, but in most instances, answers to substantive questions are "the dog," whereas the analyses that provide the bases for these answers are "the tail."

When presenting the models themselves, most readers will probably be able to understand the analyses more readily if the models are presented using the "Raudenbush and Bryk" or "HLM" style (the style used in this chapter) compared with the mixed effect (or mixed model) style in which all coefficients are included in a single equation. In the HLM style, each Level 1 coefficient has a separate equation, and different letters are used for Level 1 and Level 2 coefficients. Note that I am not suggesting the analyses need to be conducted using HLM. I am simply suggesting to use the HLM style to present the models. In my experience, more people understand more readily the structure of models when they are presented using the HLM style.

The following recommendations about reporting the results of MRCM analyses assume that the primary focus of a paper is the fixed effects, that is, the results of the significance tests of the relationships between variables. Although random error terms can be of interest, the hypotheses of most researchers will probably concern fixed effects such as those that were presented in the previous sections. Were two individual variables related (e.g., the slope representing the relationship between *Life Satisfaction* and *Individualism*), did this relationship vary across countries in some predictable way (e.g., the moderating relationship of *GDP*), and so forth.

This emphasis reflects, in part, the fact that although a significant random error term for a coefficient indicates that a coefficient varies (randomly), as discussed previously, the lack of a significant random error term does not mean that a coefficient does not vary in any way. Even without a significant random error term, it is still possible for a coefficient to vary nonrandomly, for example, when a country-level characteristic is found to be related to a within-country slope. Given this possibility, I tend to deemphasize the importance of random error terms. For most purposes, it should suffice to describe which effects were modeled as random and which effects were fixed, and if they were fixed, why. Nevertheless, this issue is far from resolved among multilevel scholars, and researchers will need to make their own decisions about the results they report.

When reporting results, it will usually suffice to describe the fixed coefficients (usually γs of some kind), the significance level associated with the test of the hypothesis that the coefficients are 0, and the *t* values. There is no need to report the coefficient, the *t* value, and the standard error because the *t* value is simply the coefficient divided by the standard error. Analysts may want to report degrees of freedom, which will indicate whether an effect was modeled as fixed or random, although the probabilities associated with *t* values do not vary as a function of degrees of freedom as they would

in an OLS analyses because in MRCM *t* values are approximations. Some other tests, such as multiparameter tests, produce chi-squares indicating differences in model fits, and for such tests, reporting the chi-squared and its associated degrees of freedom and *p* value is appropriate. Finally, in many, if not most, instances, it is probably inappropriate to report the results of the significance test of whether the intercept is significantly different from 0. Such results should be reported only when 0 is a meaningful value for the dependent measure, which, for example, would not be the case for a dependent measure that used a 1 to 7 scale.

Some authors present comparisons of models, similar to what is done when presenting SEM analyses. Although there can be good reasons to do this, for many if not most purposes, it is probably not necessary. Descriptions of MRCM analyses tend to focus on individual parameters more than the overall fit of a model, whereas in SEM, the overall fit of a model is typically the focus of the analyses. The overall fit of two MRCM models can be compared to make decisions about the inclusion of random error terms, although such procedures are probably not necessary for evaluating the inclusion or exclusion of individual error terms. The relative fits of sets of predictors can also be compared, although analysts should be certain to use full maximum likelihood estimator when doing so (e.g., Raudenbush et al., 2004).

Finally, and most important, authors are strongly encouraged to include predicted values in the interpretation of their results. For continuous variables, a commonly accepted standard is $\pm$ 1 *SD* above and below the mean. When groups are involved (at either level of analysis) it is typically helpful to generate predicted values for each group. Such predicted values can be particularly useful when explaining complex findings such as cross-level interactions.

MODEL BUILDING

Broadly speaking, there seem to be two traditions in OLS regression: one in which predetermined sets of predictors are included and retained regardless of their statistical significance, and another, more exploratory approach, in which many predictors are included and only those that are significant are retained. Certainly, other possibilities exist and are used. Nevertheless, neither of these approaches may be appropriate for MRCM, particularly at Level 1 and particularly when the number of predictors is large.

This difference in strategies reflects one of the critical differences between the OLS regression and MRCM. In OLS regression, for each predictor only

a fixed effect is estimated, and only one error term is estimated for the whole model. In contrast, in MRCM, for each Level 1 predictor, a fixed effect and a random effect are estimated, and the covariances among the random error terms are also estimated. This means that in a MRCM analysis, the number of parameters being estimated increases nonlinearly as predictors are added. For example, with a Level 1 model with no predictors, two parameters are estimated: fixed and random effects for the intercept. With one Level 1 predictor, five parameters are estimated: fixed and random effects for the intercept and the slope and the covariances between the error terms for the intercept and the slope. With three predictors, nine parameters are estimated: fixed and random effects for the intercept and the two slopes, and the covariances between the three error terms. And so on.

When thinking of building a model, statisticians will sometimes refer to the *carrying capacity* of the data. How many parameters can a data set estimate reliably? For these purposes, it may be helpful to think of a data structure as an information array and to think in terms of the amount of information that is available to estimate the parameters in a model. Assuming the same data structure, models that estimate more parameters will have less information for each parameter. At some point, a model requires too many parameters, and there is not enough information to estimate all or any the parameters.

Given this, multilevel modelers tend to favor forward-stepping algorithms, that is, adding predictors one at a time and deleting those that are not significant or testing smaller models first and adding to them. This procedure is meaningfully different from backward-stepping algorithms in which many predictors are simultaneously included at the outset, and those that are not significant are deleted. Forward-stepping approaches tend to build smaller models with fewer, but more stable, parameter estimates than backward-stepping approaches. Of course, individual analysts will have to be guided by the specific situation they face and the norms in their disciplines. Regardless, analysts who are accustomed to including simultaneously many predictors in models may have to confront the possibility that they cannot include as many (perhaps not nearly as many) predictors in their Level 1 models as they may want to include, or have been accustomed to including, in OLS regression.

USING ESTIMATED COEFFICIENTS IN OTHER ANALYSES

The coefficients estimated in a MRCM analysis can be saved and used in other analyses; however, analysts are urged to be very cautious when doing this. First, as is the case with any technique involving multiple predictors, the

estimates from any specific analysis will reflect the covariances among the variables that are included in the analysis, and with MRCM, such estimates will also reflect whatever error structure was in the model. Moreover, if a random error term is not estimated for a coefficient and no Level 2 predictor is in a model, some programs, such as HLM, will provide the same estimate for a coefficient for all Level 2 units (countries). This is a result of the algorithm that is used to estimate coefficients.

Analysts need to keep in mind that although it may be useful to think of MRCM models as series of nested equations, in fact MRCM analyses rely on a single equation that simultaneously includes predictors at all levels. Second, if shrunken estimates are used (based on a Bayesian analysis, something commonly used in MRCM programs), they are typically highly correlated with OLS estimates. They differ from OLS estimates in terms of their variances, and so for pure correlational analyses, there may not be much a difference in results using the two. Third, unless the coefficients are used in an analysis in which random error can be modeled, the analysis will not reflect this aspect of the multilevel analysis, undermining the value of using MRCM to estimate the coefficients.

With this in mind, analysts are encouraged to think creatively about how they can examine questions of interest within the multilevel context. For example, within-country differences such as changes across time can be examined using various coding schemes that reflect the trends of interest. Groups of countries can be compared by creating variables at Level 2. Analysts should think carefully before using the coefficients estimated by MRCM in some other type of analysis.

WEIGHTS

Decisions about weights in analyses are not unique to multilevel modeling. For various reasons, researchers may want to assign more importance (more weight) to some observations than to others. Such decisions frequently reflect desires to correct samples for the under- or overrepresentation of different types of respondents. For example, if 75% of the respondents in a sample are men and it is known that the population has the same number of men and women, each man might be weighted 2 and each women 3 (or .67 and 1, or some ratio reflecting the desired adjustment). There are no hard-and-fast rules for weighting observations, and such adjustments can reflect various considerations. See Haedler and Gabler (2003) for a broad discussion of weighting, and see the European Social Survey (ESS, 2009) for an excellent discussion of the use of weights within a multilevel, cross-cultural study.

The aspect of weighting that is unique to multilevel modeling is the fact that weights can be assigned to units at each level of analysis (country and person for present purposes). At the person level, such weights are sometimes referred to as *design weights* because they reflect the distribution of cases obtained from a certain sampling design. Design weights can be used to adjust for the lack of representativeness in a sample. Such nonrepresentativeness may be intentional. For example, certain groups may be oversampled (data are collected from proportionately more people than exist in the population) to ensure that researchers have adequate samples for groups with few members (e.g., immigrants). Nonrepresentativeness may also be unintentional. For example, attempts to create representative samples may fall short for various reasons. Regardless of the reasons, if samples are not representative, estimates of country-level statistics (within a multilevel framework or not) may be inaccurate because the sample on which they are based does not represent the population the statistics are meant to describe. Such problems can be addressed with appropriate weights for analyses.

At the country level, weights will usually reflect differences in the population of countries in a sample. For example, assume a study has 100 observations from both Finland and China. Between 5 and 6 million people live in Finland, whereas over 1 billion people live in China. When making inferences about humankind, it does not make much sense to treat these two groups of 100 observations similarly. Exactly how to take into account such disparities may not be fixed. A researcher may weigh countries (Level 2 observations) using population per se, log or other transforms of population, and so forth.

DESIGN CONSIDERATIONS AND ALTERNATIVE ANALYTIC STRATEGIES

Similar to the consideration of single-level data structures, researchers frequently have questions about the power of multilevel data design. How well can a sample of such-and-such detect a difference of such-and-such? Unfortunately, estimating the power of multilevel models is not well-understood, and the following discussion should be interpreted as a discussion of guidelines rather than a description of hard and fast rules. Various rules, usually described in terms of the number of Level 2 and Level 1 observations, have been proposed such as, 30/30 (Kreft, 1996), 50/20 or 100/10 (Hox, 1998). It should be noted that these rules reflect research designs in which obtaining more Level 2 units is probably much easier than it is for cross-cultural

researchers to obtain more countries. See Richter (2006) for an informed and accessible discussion of power analyses in multilevel models.

As is the case with single-level analyses, the axiom "more is better" certainly applies. More observations (at either level of analysis) will provide more stable parameter estimates. Unfortunately, particularly for cross-cultural research, it may be difficult, if not impossible, to obtain more observations, particularly at the country or culture level. Moreover, given the difficulty of obtaining data from multiple countries, researchers are often faced with am empirical fait accompli – country-level data have been collected, and the essential question is "Are there enough countries to justify a multilevel analysis?"

Answering this broader question requires answering other questions, the most important of which is probably this: Do the primary hypotheses concern intercepts or slopes from Level 1? Intercepts are invariably more reliable than slopes, and so it is easier (i.e., the data provide more power) to examine country-level differences in intercept than in slopes. More reliable coefficients are more efficient, meaning that fewer countries will be needed to model country-level differences in intercepts compared with slopes. Nevertheless, many hypotheses concern country-level differences in slopes, for example, how does the relationship between individual-level variables X and Y vary as a function of country-level variable Z?

Given that many cross-cultural researchers will be interested in country-level differences in within-country coefficients (intercepts and slopes), this discussion focuses on the number of countries that are needed. First, it is informative to think of the countries in an analysis as being sampled from the population of countries. Within such a context, analysts can think of the minimum number of observations they would need in a traditional, single-level design. Certainly, few researchers would consider a study in which there were only 4, 5, or 6 observations. Similarly, within the multilevel context, a small number of countries will not provide a good basis for estimating parameters, and such shortcomings will be particularly important for estimating random error terms. Coefficients that theoretically should be random will not be able to be modeled as such because there will not be enough information to estimate the random error component.

All this is well and good but begs the question, What can or should an analyst do if there are simply not enough countries to perform a multilevel analysis, for example, there are only two or three countries? Various options are available in such cases. If the research questions primarily concern differences in means, then some type of analysis of variance may be appropriate in which countries are treated as a between-groups factor. If

the research questions primarily concern within-country relationships, the most appropriate option is probably a technique known as *regression by groups*. As mentioned in the introduction to this chapter, in such analyses, separate regression equations are estimated for each group (country), and then the coefficients for each group can be compared. Although this technique does not model the random variability in coefficients, it does not assume that coefficients (slopes) are equal across groups. Finally, I urge analysts to be cautious regarding the use of LSDV analyses. As explained in the introduction, unless such analyses explicitly include terms that test the similarity of slopes across groups (countries), the analyses assume that the slopes are similar.

It is important to keep in mind however, that such analyses do not provide the opportunity to model (to explain statistically) between-country differences in coefficients (either means or slopes). Although they may be able to establish the fact that coefficients vary across countries, they do not provide a statistical basis for drawing conclusions about why such differences exist. There are simply not enough countries to provide a basis for statistical inference. In such cases, analysts will need to rely on other bases for drawing conclusions about why countries differ.

SOFTWARE OPTIONS

In response to the growing interest in multilevel modeling, the number of programs that can perform such analyses has increased meaningfully since 1995. There are too many programs to discuss specific comparisons here, so I offer only broad guidelines. First, it is important to note that the major programs (i.e., programs offered by reputable software firms) will give the same results, *provided the same models are specified*. I have emphasized *the same models* because multilevel models require analysts to make numerous decisions about error structures, centering, and so forth, and different programs will implement different options in different ways.

For analysts who are familiar with the subtleties of MRCM, the selection of software can be guided by familiarity and accessibility. The situation is different for analysts who are not familiar with MRCM. For such analysts, I recommend programs such as HLM and MlwiN that have been specifically designed to do MRCM analyses, with HLM probably being the easiest to use. This recommendation is based on the ease with which models can be specified (error terms, centering, etc.), the ease with which the output can be interpreted, and my experience giving workshops. For analysts who are not familiar with MRCM, the array of options in many multipurpose programs

can be confusing and can lead them to specify (unwittingly) inappropriate or incorrect models because such programs were designed to conduct more than just MRCM.

Finally, the availability of different software options highlights the importance of referring to the analyses discussed in this chapter as multilevel random coefficient models or sometimes simply as multilevel models (MLM) instead of using the term hierarchical linear models . In the multilevel world, HLM is a specific program that conducts MRCM, and authors should be careful to distinguish the statistical technique they used (MLM) from the program they used to conduct the analyses (e.g., HLM, SAS, etc.).

CONCLUDING REMARKS

The fit between multilevel modeling and cross-cultural research is a natural one, particularly for cross-cultural psychologists who are invariably interested in the joint effects of individual and cultural differences on the attitudes, beliefs, and behaviors of individuals. Such interests typically rely on analyses of data sets in which observations from multiple individuals within multiple cultures are collected, and multilevel modeling provides a comprehensive framework within which various types of relationships can be examined within such data structures.

My sense is that because of this fit, the use of multilevel modeling in cross-cultural research will grow over the next few years and beyond. Such growth will not be without its growing pains, however. Particularly for researchers whose primary training has concerned single-level OLS techniques, the considerations involved in conducting multilevel analyses can be daunting. Coefficients at multiple levels must be interpreted simultaneously, and relationships at different levels of analysis influence each other. Moreover, there are the complexities of modeling error structures.

One of the obstacles to the widespread use of multilevel modeling is the relative lack of university courses on the topic. Although this is changing, by no means is multilevel modeling currently part of the accepted canon for graduate education in statistics in the social sciences. Nevertheless, there are excellent resources available, in terms of books and articles, and professional training (e.g., workshops), and interested researchers can develop an understanding of multilevel modeling well after they have finished their formal training. One of my goals in writing this chapter was to provide enough information to allow researchers who are unfamiliar with the technique to consider using MRCM, or at the least provide them with enough information to understand the results of studies using MRCM. The longest journey begins with a single step.

REFERENCES

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

Baron, R. M., & Kenny. D. A. (1986). The moderator–mediator distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1182.

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods, 11*, 142–163.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Newbury Park, CA: Sage.

European Social Survey. (2009). Weighting European social survey data. London: Centre for Comparative Social Surveys, City University, London.

Haedler, S., & Gabler, S. (2003). Sampling and estimation. In J. A., Harkness, F. J. R. Van de Vijver, & P. Ph. Mohler (Eds.), *Cross-cultural survey methods* (pp. 117–134). New York: Wiley.

Hox, J. J. (1998). Multilevel modeling: When and why? In I. Balderjahn, R. Mather, & M. Schader (Eds.), *Classification, data analysis and data highways* (pp. 147–154). New York: Springer.

Kenny, D. A., Korchmaros, J. D., & Bolger, N. (2003). Lower level mediation in multilevel models. *Psychological Methods, 8*, 115–128.

Krull, J. L., & MacKinnon, D. P. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate Behavioral Research, 36*, 249–277.

Kreft, I. G. G. (1996). *Are multilevel techniques necessary? An overview, including simulation studies*. Retrieved March 31, 2006, from http://www.calstatela.edu/faculty/ikreft/quarterly/quarterly.html.

Kreft, I. G. G., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Newbury Park, CA: Sage.

Marsh, H. W., & Hau, K. (2003). Big-fish-little-pond-effect on academic self-concept. A cross-cultural (26 country) test of the negative effects of academically selective schools. *American Psychologist, 58*, 364–376.

Nezlek, J. B. (2001). Multilevel random coefficient analyses of event and interval contingent data in social and personality psychology research. *Personality and Social Psychology Bulletin, 27*, 771–785.

Nezlek, J. B. (2003). Using multilevel random coefficient modeling to analyze social interaction diary data. *Journal of Social and Personal Relationships, 20*, 437–469.

Nezlek, J. B., & Plesko, R. M. (2003). Affect- and self-based models of relationships between daily events and daily well-being. *Personality and Social Psychology Bulletin, 29*, 584–596.

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interaction effects in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics, 31*, 437–448.

Rabash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., et al. (2000). *MLn: Command reference guide*. London: Institute of Education.

Raudensbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models* (2nd ed.). Newbury Park, CA: Sage.

Raudenbush, S., Bryk, A., Cheong, Y. F., & Congdon, R. (2004). *HLM 6: Hierarchical linear and nonlinear modeling.* Lincolnwood, IL: Scientific Software International.

Richter, T. (2006). What is wrong with ANOVA and multiple regression? Analyzing sentence reading times with hierarchical linear models. *Discourse Processes, 41*, 221–250.

Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics, 23*, 323–355.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis.* London: Sage.

APPENDIX

Sample data set Leve1 1 data

Coun.	Life	Ind.	Ind2.	Union	NUn.	Ucnt.	Coun	Life	Ind.	Ind2.	Union	NUn.	Ucnt.
A	1	1	1	0	1	−1	D	5	5	5	1	0	1
A	2	2	2	1	0	1	D	6	6	6	1	0	1
A	3	2	2	1	0	1	D	6	7	7	1	0	1
A	4	3	3	1	0	1	D	7	8	8	1	0	1
A	5	3	3	0	1	−1	D	8	9	9	1	0	1
A	6	3	3	1	0	1	D	9	10	10	1	0	1
A	7	4	4	1	0	1	E	1	4	104	0	1	−1
A	8	4	4	1	0	1	E	2	5	105	1	0	1
B	2	2	2	0	1	−1	E	3	6	106	0	1	−1
B	2	2	2	0	1	−1	E	4	6	106	0	1	−1
B	3	2	2	0	1	−1	E	5	6	106	1	0	1
B	3	4	4	0	1	−1	E	6	7	107	0	1	−1
B	4	4	4	1	0	1	E	7	8	108	1	0	1
B	4	4	4	1	0	1	E	8	9	109	1	0	1
B	5	5	5	1	0	1	E	9	10	110	0	1	−1
B	6	5	5	0	1	−1	E	9	10	110	0	1	−1
B	6	5	5	0	1	−1	F	1	1	101	0	1	−1
C	3	3	3	0	1	−1	F	1	2	102	0	1	−1
C	3	3	3	1	0	1	F	4	5	105	0	1	−1
C	3	3	3	0	1	−1	F	5	7	107	1	0	1
C	4	6	6	1	0	1	F	5	8	108	1	0	1
C	4	6	6	1	0	1	F	6	10	110	0	1	−1
C	4	6	6	1	0	1	G	5	2	102	0	1	−1
C	5	8	8	1	0	1	G	5	2	102	0	1	−1
C	5	8	8	0	1	−1	G	5	2	102	0	1	−1
C	5	9	9	1	0	1	G	6	3	103	0	1	−1
C	6	9	9	1	0	1	G	6	3	103	1	0	1
D	2	1	1	0	1	−1	G	6	3	103	1	0	1
D	3	2	2	1	0	1	G	7	4	104	1	0	1
D	4	3	3	1	0	1	G	7	4	104	1	0	1
D	4	4	4	0	1	−1	G	7	5	105	0	1	−1
G	8	6	106	0	1	−1	I	6	6	106	0	1	−1

Coun.	Life	Ind.	Ind2.	Union	NUn.	Ucnt.	Coun	Life	Ind.	Ind2.	Union	NUn.	Ucnt.
H	4	1	101	0	1	−1	I	7	7	107	1	0	1
H	4	1	101	0	1	−1	I	7	7	107	1	0	1
H	4	2	102	1	0	1	I	7	8	108	1	0	1
H	5	2	102	1	0	1	I	7	9	109	1	0	1
H	5	3	103	1	0	1	I	8	10	110	0	1	−1
H	5	4	104	0	1	−1	I	8	10	110	0	1	−1
H	6	5	105	1	0	1	J	3	4	104	0	1	−1
H	6	5	105	1	0	1	J	3	4	104	0	1	−1
H	6	6	106	1	0	1	J	3	5	105	1	0	1
H	6	7	107	0	1	−1	J	4	5	105	1	0	1
H	7	5	105	1	0	1	J	4	5	105	1	0	1
H	7	5	105	1	0	1	J	5	8	108	0	1	−1
I	6	4	104	0	1	−1	J	5	8	108	0	1	−1
I	6	4	104	0	1	−1	J	7	8	108	0	1	−1
I	6	5	105	0	1	−1	J	7	8	108	0	1	−1

Coun. – Country identifier
Life – Life satisfaction
Ind. – Individualism
Ind2. – Individualism with 100 added to observations for some countries
Union – Dummy-coded variable representing membership in a union, member = 1, non-member = 0.
NUn. – Dummy-coded variable representing non-membership in a union, nonmember = 1, member = 0.
Ucnt. – Contrast-coded variable representing membership in a union, member = 1, non-member = −1.

Sample data set Level 2 data

	Means of Level 1			
Country	GDP	Individ	Individ2	Party
A	8	2.75	2.75	0
B	12	3.67	3.67	0
C	14	6.10	6.10	0
D	9	5.50	5.50	0
E	12	7.10	107.10	0
F	14	6.00	106.00	1
G	12	3.40	103.40	1
H	15	3.83	103.83	1
I	15	7.00	107.00	1
J	16	6.11	106.11	1

Country – Country identifier
GDP – Gross Domestic Product
Individ. – Country level mean of level 1 variable Ind
Individ2. – Country level mean of level 1 variable Ind2
Party – Dummy-coded variable representing multiparty political system in a country, 0 = dual-party system and 1 = multiparty system