5 Strategy for Measuring Constructs and Testing Relationships

In this chapter, we discuss the statistical methods used to test the viability of our conceptual models as well as the methods used to test our hypotheses. We first discuss the justification for aggregating our measures to the CEO level of analysis. Next, we present the strategies used to assess the psychometric properties of our scales. We include a discussion of the critical issue of measurement equivalence that assesses whether our scales were interpreted similarity across countries. In this chapter, we also describe a new measure, Gestalt Fit, which we used to test our theoretical proposition that the match between country-level leadership expectations (culturally endorsed implicit leadership theory, or CLT) and actual CEO behavior leads to critical outcomes. This new measure assesses the fit between CLTs and CEO behavior across all dimensions and is theoretically driven from and consistent with the GLOBE conceptual model presented in Chapter 1.

Finally, we also discuss the rationale for the particular statistical techniques used to test our hypotheses. In order for hypotheses to be fairly and accurately tested, we have to be confident that the statistical techniques were used appropriately and did not introduce statistical conclusion errors (Hanges & Wang, 2012). These errors occur when the empirical findings lead researchers to false conclusions and are caused by inappropriate choice of statistical analysis or violations of statistical assumptions required for a particular statistical technique (e.g., violation of the independence of errors and assumption observations). Specifically, we used hierarchical linear modeling (HLM) to test the GLOBE hypotheses because of the nested nature of the data.

There are several statistical techniques that we used to help us test the viability of our conceptual models and hypotheses. For example, we hypothesized that members of a top management team (TMT) would agree in their description of the leadership behaviors of their CEO. That is, we hypothesized that because of sufficient shared experience with the CEO, the average leader behavior rating would be more reflective of the CEO than the individual perceptions of the TMT members. How do we demonstrate this? We will discuss that in this chapter.

Next, we discuss the evidence that we collected to assess the extent to which we have valid scales measuring each construct. For our purposes, this involves demonstrating not only that our scales are internally consistent within a country but that they are equivalent across various countries. Thus, we will discuss the methodology that we developed and used to assess cross-culture measurement equivalence.

Following this discussion, we present the methodology used to test the hypotheses concerning the relationships among culture, leadership expectations (i.e., CLTs), and leadership behavior. Subsequent to this, we present the methodology used to test the effectiveness of specific leadership behaviors such as Charismatic leadership. Following this discussion, we describe our fit hypotheses that reflect our belief regarding the importance of the match between the CEO behavior and the country-specific CLTs. Finally, we will review the methodology used to separate particularly effective from particularly ineffective CEOs.

It should be noted that the purpose of this chapter is only to describe the logic and relevant mathematical aspects underlying these statistical procedures. We leave the presentation, interpretation, and discussion of results to the subsequent chapters in this book. The intent of this chapter is to provide a framework to aid the reader's understanding and interpretation of the subsequent chapters.

Aggregation

The constructs that we are measuring in this study are what Kozlowski and Klein (2000) called convergent-emergent constructs. Our constructs start as a function of the individual psychological processes (i.e., cognition, affect) of our TMT respondents. However, even though the responses are a function of the individual TMT's psychological processes, their responses *converge* to a single value, represented by an average. This convergence occurs because the scale items are assessing some shared reality (e.g., employees' experience of organizational culture or TMT members' interactions with the CEO). For example, TMT members have their own unique impression of the CEO's participative leadership behavior. However, the TMT members' perceptions tend to converge because they are describing the same external reality, namely, the CEO's participative behavior. In addition to being convergent, the constructs are said to be *emergent* because the psychometric properties of these scales only operate or are only valid at the aggregate- or group-level of analysis (Hanges & Dickson, 2004). For instance, one can only have a valid discussion of Team Solidarity by measuring this variable at the team level of analysis.

Given that we hypothesized that TMT members' responses to our items supposedly measuring a particular construct should converge, we need to demonstrate that this is actually occurring. Similar to our prior quantitative work (House, Hanges, Javidan, Dorfman, & Gupta, 2004), we mainly relied on the intraclass correlation coefficients (ICC[1]) and one-way analyses of variance (Bliese, 2000; Shrout & Fleiss, 1979) to demonstrate that convergence was occurring, and thus, aggregation of individual responses to the mean response for a construct was justified. Median values of ICC(1) of .12 are reported in the organizational literature (James, 1982). As will be presented in Chapter 6, our ICCs for each scale compare very favorably to the literature. We also computed $r_{WG(J)}$ (James, Demaree, & Wolf, 1984) to demonstrate that aggregation of individual scores was justified. The $r_{WG(J)}$ analysis is an index that assesses the degree to which there is within-group agreement. Finally, we also examined the reliability of the TMT managers' average estimate (ICC[2]) for each scale. Our findings were very good for each scale thereby indicating that aggregation was appropriate. The $r_{WG(J)}$ analyses are reported for each scale and are presented in Chapter 6.

Creating Psychometrically Sound Scales

The following procedures were used to create our scales. First, we aggregated the responses from the TMT members to the CEO/organization level of analysis. Recall that the CEO and the organization level of analysis are equivalent because there is one to one correspondence for CEOs and organizations. TMT responses had to be aggregated to develop our scales because no TMT member responded to every leadership item for a particular scale. As discussed in Chapter 4, items for leadership dimensions (e.g., Charismatic: visionary) were separated across Surveys A and B. Each TMT member only completed one of these surveys. Thus, the covariance between the items on Survey A and those items on Survey B could not be estimated at the individual TMT member level of analysis. This covariance could only be estimated at the CEO/organization level of analysis and that is why we used aggregated data for this analysis and for all subsequent analyses. That is, because of the choice in our research design minimize same source error, individual TMT respondents only completed selected items for our constructs—thus, the necessary psychometric analyses for our scales could not be performed at the individual TMT respondent level.

We then standardized the average item responses within each country to remove country-level differences. Country-level differences are eliminated by this standardization procedure because the standardized scores for each scale will have a mean of zero within each country. Subsequent to this standardization, we performed a maximum likelihood exploratory factor analysis for each scale on the data pooled across all countries. This exploratory factor analysis is called a pooled within-country factor analysis and reflects the factor structure of the scales at the CEO/organization level of analysis across all countries. Recall that the CEO and organizational level of analysis are identical because there is only one CEO per organization. We also computed Cronbach's coefficient alpha to assess the degree of internal consistency reliability for these factors. The results of these analyses are also presented in Chapter 6.

Measurement Equivalence

While the prior factor analysis established the average within-country factor structure of our scales, we also need to establish the measurement equivalence of these scales prior to testing our GLOBE hypotheses. That is, we need to demonstrate that a similar factor structure actually exists in every country. Measurement equivalence can exist at several levels, and one does not have to attain the highest level in order to use a particular scale (Steenkamp & Baumgartner, 1998).

The first and lowest level of equivalence is called *construct equivalence,* and it is demonstrated by showing that the same number of factors emerges among a set of items and that the same items "load" on the same factors across countries. The second and next level of equivalence is called *metric or measurement unit equivalence*. Metric equivalence is demonstrated by determining whether the factor loadings are equivalent across countries. If this level of equivalence is demonstrated, there is evidence that the scale has equal scale intervals across countries and thus, assessment of statistical relationships (e.g., correlation, regression) is meaningful. The third and next level of measurement equivalence is *scalar equivalence*. This level assesses whether the items are either upwardly or downwardly biased between countries. That is, the true score for two countries on a particular scale (e.g., Team-Oriented leadership) could be truly equal but because of response bias, the average scale scores for these two countries would be different (if the scale lacked scalar equivalence). Scalar equivalence is assessed by examining the equivalence of the intercepts across countries. This level of equivalence is important for meaningful comparison of scale means or averages across countries.

The fourth and final level is labeled *full equivalence* in which the equivalence of the latent variance/covariance matrix and the error variance/covariance matrix across countries are examined. If the error variance/covariance matrix is equivalent across countries, then that means that the construct is measured with equal reliability across countries. This final level of measurement equivalence is an extremely difficult criterion and rarely obtained but provides a more complete understanding of the extent to which country-level variables have little or no effect on responses to survey items.

For the present project, we are primarily interested in assessing relationships among constructs. In order to meaningfully interpret these relationships among constructs (e.g., Participative leadership and Firm Competitive Performance), it is important that the metric level of equivalence exists for our scales. For example, participative leadership may relate to certain outcomes in the United States such as higher motivation for TMTs under a participative CEO, but not in Japan. This could be due to people interpreting the scale of participation differently in the United States and Japan—a relatively uninteresting finding. Conversely, the different relationships may be due to the fact that people in both cultures are interpreting the scale the same, but the effects of participatory leadership are different across countries. This latter finding is more theoretically and practically interesting.

Usually all four levels of measurement equivalence are assessed by using structural equation modeling. In particular, a multigroup structural equation model is performed in which each country's data is used and progressively more constraints are placed on the model as the lower levels of measurement equivalence are met. Minimal sample size for these types of analyses is 100 observations per group. Unfortunately we could not use this approach because our sample size was approximately 40 observations per country. Therefore, we developed an alternative protocol. This protocol was based on a meta-analytic perspective and assessed the extent to which our data had both construct and metric equivalence.

To demonstrate the first level of measurement equivalence (i.e., construct equivalence), we first conducted principal components analysis separately in each country and examined the eigenvalues for each scale. We computed the average eigenvalue across countries and computed a 95% confidence interval across countries for the extracted factors. Construct equivalence was declared if the lower bound of the confidence interval for a particular factor exceeded the eigenvalue > 1 rule. Therefore, the number of factors extracted for a set of items was determined by the number of factors whose lower bound of the confidence interval was higher than 1. In almost all cases, a single factor was adequate to represent the scale thus demonstrating the first level of measurement equivalence—construct equivalence. For example, in Table 6.6, the first factor for the visionary scale (Survey C items) had an eigenvalue that ranged from 2.91 to 4.62 across the 24 countries in our sample. That means that, across the 24 countries, this single factor explained anywhere from 48.5% to 77% of the original item variance. In other words, a single factor was extracted in all of our countries (i.e., construct equivalence) for this scale and that this single factor did a very good job summarizing the single visionary leadership construct among these items in all countries.

Metric Equivalence

To demonstrate metric equivalence, we conducted a series of within country exploratory factor analyses. It should be noted that when performing exploratory analyses, the resultant factor structure is a function of the population factor structure as well as some unique characteristics of the sample. Thus, when initially examining the factor structure across two countries, it might appear that there are meaningful differences in the factor structure across countries. Therefore, after obtaining each country's factor structure, we performed a Procrustes rotation to compare the similarity of each country's factor loadings. Procrustes rotation compares the factor structure of a pair of countries (or the factor structure between one country and some comparison sample's factor structure) and determines the degree to which the two country factor structures can be rotated to achieve total similarity. In other words, Procrustes rotation seeks to determine the extent to which the country differences in factor structure are more apparent than real.

The procedure to compute a Procrustes rotation was as follows. We first created a comparison sample that we used to compare the similarity of each country's factor structure for a particular scale. We used the pooled within-country data set, described earlier, as our comparison sample and obtained the factor loadings from this data set. Following the work of other cross-cultural researchers, the pooled within-country data set was created by obtaining the correlation matrix for a set of items within each country, converting these correlations to the Fisher Z equivalents, averaging the transformed correlations across countries and then reconverting these average values back into correlations to obtain a pooled within-country correlation matrix. We then performed a factor analysis on this pooled within-country correlation matrix, and the resultant factor structure was used as the comparison structure for all of our Procrustes rotations for a particular scale.

The specific results of this comparison for each scale are presented in Chapter 6. In general, our findings supported the metric equivalence of our scales—that is, factor loadings were equivalent across countries. This indicates that the relationships among our variables can be meaningfully interpreted across countries.

Statistical Analyses Testing GLOBE Hypotheses

As indicated in Chapter 4, we collected data from several different sources and across multiple levels of analysis in this study. Specifically, TMT members (Level 1: Individual TMT members) provided information about the CEO's leadership behavior as well as the organizational performance (Level 2: CEO/organization level of analysis). We sampled an average of 40 organizations per country with TMT members only appearing in our sample for a single organization (i.e., TMT members were nested within organizations). Further, these organizations were nested within countries (Level 3: Country). As indicated previously, we used the country-level CLT information presented in House and colleagues (2004) to provide information about leadership expectations for each country. Thus, statistical analyses used in this study were chosen in order to handle the nested data structure.

Research designs that have variables at multiple levels of analysis have been referred to as multilevel (Kozlowski & Klein, 2000), cross-level (Rousseau, 1985), meso (House, Rousseau, & Thomas, 1995), or mixed-determinant (Klein, Dansereau, & Hall, 1994) models or theories. As previously indicated, our research design and our hypotheses are multilevel and therefore require appropriate statistical analysis for this kind of data. We tested many of our hypotheses using a technique known to be an effective tool for analyzing multilevel conceptual models and nested data—HLM (Hofmann, 1997; Hofmann, Griffin, & Gavin, 2000). HLM, also referred to as multilevel linear models in the sociological research (Goldstein, 1995), mixed effects and random effects models in the biometrics literature, random coefficient regression models in the econometrics literature, and as covariance components models in the statistical literature (Bryk & Raudenbush, 1992). We choose this analysis because of the nature of our variables, the nature of our hypotheses, and the structure of our data. We should note that due to the nature of our research design, we could not run a three level HLM analysis. As discussed earlier, the data at Level 1 (i.e., TMT member level) were averaged to the CEO/Organization level of analysis before our analyses. Thus, we performed a 2 level HLM (CEO/Organization level and Country level).

HLM can be conceptualized as a multistep process designed to test relationships between independent and dependent variables at multiple levels. The following example is used to explain how the analysis was conducted. One hypothesis that was tested was that CEO leadership behavior (e.g., Charismatic) would predict TMT Effort and Firm Competitive Performance. This involves multilevel analysis because even though the independent variable (CEO leadership behavior) and the dependent variables (TMT Effort and Firm Competitive Performance) are at the same level of analysis (i.e., Level 1: CEO/organization), we need to account for the fact that the organizational data is nested within countries (i.e., Level 2: Country).

The first step of our HLM analysis can be thought of as producing an equation for each country between CEO Charismatic leadership and one of our dependent measures (e.g., TMT Effort). Equation 5.1 shows the Level 1 (i.e., within-country) equation regressing TMT Effort onto Charismatic leadership.

$$\text{TMT Effort}_{ij} = \beta_{oj} + \beta_{1j}\ \text{Charisma}_{ij} + r_{ij} \tag{5.1}$$

In this equation, β_{oj} refers to the intercept for country j. It represents the unadjusted mean TMT effort in country j. In Equation 5.1 above, β_{1j} represents the unstandardized slope for the relationship between Charismatic leadership and TMT effort in country j. Finally, r_{ij} represents the error associated with estimating this equation.

It is possible that Equation 5.1 varies across countries. For example, perhaps the unadjusted mean TMT Effort might be higher in some countries (e.g., Peru, Austria) than another (e.g., Russia, Estonia). It is also possible that the slopes of Charismatic leadership–TMT Effort relationship might vary across countries. In other words, perhaps this relationship is stronger in some countries than in others. To test these possibilities, one has to conduct a Level 2 (between country) analysis.

To test whether there are significant differences in the equation 5.1 regression coefficients, the following two equations would be computed:

$$\beta_{oj} = \gamma_{oo} + U_{oj} \tag{5.2}$$

$$\beta_{1j} = \gamma_{10} + U_{1j} \tag{5.3}$$

In equation 5.2, γ_{00} represents the grand intercept averaged across all countries and U_{oj} represents the between country variability among the intercepts. There is a χ^2 test that assesses whether U_{oj} is significantly larger than would be expected if there were no real differences in intercepts among the countries. This type of analysis is referred to as a random intercepts model (Bryk & Raudenbush, 1992; Kreft & Leeuw, 1998). In random intercept models, only the means of the dependent variable (i.e., the intercept in equation 5.1) are allowed to vary across countries and the focus of such analyses are usually to predict why this country-level variation is occurring.

In Equation 5.3, γ_{10} represents the grand slope averaged across all countries. U_{1j} represents the between country variability among the slopes. As with the random intercept model, there is a χ^2 test that assesses whether U_{1j} is significantly larger than would be expected if there were no real slope differences among the countries. This kind of analysis is called a random slopes model (Bryk & Raudenbush, 1992; Kreft & Leeuw, 1998). In random slope models, if there are significant country-level differences in the slope, some country-level variable is identified to help explain why the slopes vary. Basically, random slopes models can roughly be thought of as similar to traditional moderated multiple regression analysis in which some variable (e.g., culture) is believed to moderate the relationship between two other variables (e.g., Charismatic leadership–TMT Effort relationship).

Before closing this section, there are two final points. First, we used grand-mean centering in our HLM analysis. We choose grand-mean centering in this study because it enables more meaningful tests of interactions (Kreft & Leeuw, 1998) as well as tests for incremental variance of subsequent predictors in the analysis. Second, when we report percentage of variance explained (R^2) from our HLM analyses, we are reporting R^2s for the specific level at which the analysis was being conducted. Therefore, if we tested a predictor at Level 1, the R^2 associated with that predictor is the percentage of Level 1 variance explained by that predictor.

Assessment of Culturally Endorsed Implicit Leadership Theory–Behavior Fit

As shown in Figure 1.1, one of the major hypotheses in this study is that the fit between culturally endorsed implicit leadership theory (CLT) and CEO leadership behavior affects leader acceptance and effectiveness. The question that will be addressed in this section is how to assess fit so that this hypothesis can be empirically verified. We first considered using the response surface methodology recommended by Edwards (1995, 2002). However, it became clear that this approach would focus on each leadership dimension separately, and thus, it is not consistent with our current conceptual understanding of how schemas and CLTs actually operate (i.e., in a gestalt manner; Hanges, Dorfman, Shteynberg, & Bates, 2006).[1] As these new cognitive models suggest, when people think of their ideal or prototypical leaders, an entire picture emerges as opposed to a dimension-by-dimension conceptualization. Thus, we develop a new fit index that is consistent with this current and more Gestalt approach to cognitive thinking.

To test our hypothesis we developed a new fit index. Our definition of *fit* is formally defined as the degree of similarity between a CEO's leadership behavior and the country-level leadership expectations (CLT scores) across the 21 primary leadership dimensions. This definition is operationalized by capturing two aspects of the match between CEO leader behavior and CLTs. The first aspect, hereafter referred to as profile pattern similarity, assesses the linear pattern between individual CEO's leadership profile and the CEO's cultural leadership profile (across the 21 leadership dimensions). The second aspect, hereafter referred to as the absolute behavioral match, assesses the overall similarity in the absolute level or magnitude between each CEO's behavior and the CLT dimensions (across the 21 leadership dimensions). These two aspects of fit, profile pattern similarity, and absolute behavioral match, were combined into a single fit index. The advantage of using this new method over previous methods is that it is consistent with current cognitive models and directly incorporates culture into the statistical analysis in a fashion consistent with our conceptual model.

Regression analyses were performed by first reformatting the GLOBE database so that for each CEO the reformatted data matrix consisted of 21 rows and 6 columns. The columns of this converted matrix represented variables specifying (1) the leadership dimension contained in each row of the data, (2) a country code variable, (3) the CLT variable ratings, the CEO leadership behavior variable, and (4) a CEO identification variable.

Because the fit index is new, we provide the following example to aid the reader's understanding of how it was computed. Table 5.1 presents simulated data for both country-level CLTs and CEO behaviors for five primary leadership behaviors. As can be seen in this table, CEO behaviors and the country specific CLTs are presented in each row for a single leadership dimension.

[1]We considered using the Edwards (1995, 2002) response surface methodology to test the fit between the CLT and CEO behavior. However, it soon became clear that the Edwards procedure was inconsistent with the underlying theory driving our project. Specifically, "fit" in Edwards' procedure is assessed element by element and would involve interaction terms between each of the 21 primary leadership dimensions with their respective country-level CLT dimensions. For example, for the Autonomous leadership scale, Edwards procedure requires the use of five predictors: autonomous behavior (x1), autonomous CLT (x2), autonomous behavior squared ($x1^2$), autonomous CLT squared ($x2^2$), and the interaction between autonomous behavior and CLT (x1x2). This procedure is then repeated for each of the remaining 20 primary leadership dimensions. Rather, as discussed in Chapter 9, we believe that the theoretical mechanism is a gestalt matching of all 21 leadership dimensions to the CLTs rather than a dimension-to-dimension analysis. Furthermore, using this procedure would require 105 predictors for each of the eight dependent variables (i.e., five predictors for each of the 21 primary leadership dimensions). A meaningful understanding of the results of such analyses, along with a visual depiction, would be virtually impossible.

Table 5.1 Illustrative Database Demonstrating Calculation of Profile Fit

	Country	CLT 2004	CEO Behavior	CEO ID
Admin. competence	1	6	5	A
Autocratic	1	5	3	A
Autonomous	1	4	2	A
Visionary	1	6	4	A
Inspirational	1	5	3	A
Admin. competence	1	6	5	B
Autocratic	1	5	3	B
Autonomous	1	4	3	B
Visionary	1	6	6	B
Inspirational	1	5	7	B
Admin. competence	2	2	2	C
Autocratic	2	1	2	C
Autonomous	2	2	3	C
Visionary	2	5	4	C
Inspirational	2	5	4	C

(While not shown in Table 5.1, in actuality, there were 21 rows of data for each CEO's data where each row provides the corresponding ratings for the 21 first-order GLOBE leadership scales.) Our example continues by considering a single country (labeled as country 1). For each primary leadership dimension such as "administratively competent," Table 5.1 presents this country's CLT score, a single CEO's rating on this dimension, and the identification of this CEO (e.g., ID for the first CEO in country 1 is labeled as CEO A). The first 5 rows of this table represent data for this specific CEO in country 1 (i.e., CEO A). The second five rows represent CEO B for the same five leadership behaviors. The last five rows show the data for CEO C for the same leadership behaviors along with the CLT for another country (denoted as country 2 in the first column in Table 5.1).

Pattern Similarity Fit

To assess the pattern similarity fit between each CEO's leadership behavioral profile and the CLT leadership profile, we conducted separate simple linear regressions for each CEO using the House and colleagues' (2004) 21 CLT leadership dimensions as the predictor and the measured (i.e., actual) 21 CEO leadership behaviors for the same dimensions as the dependent variable. The unstandardized regression coefficient for the slope between these variables was our measure of pattern similarity fit between the CLT leadership dimensions and the CEO behavioral dimensions. In other words, the following unstandardized regression equation was computed separately for *each* CEO:

$$Behavior = b_{0i} + b_{y_i.x_i} CLT \tag{5.4}$$

In this equation, $b_{y_i.x_i}$ represents the unstandardized slope between leadership CLT and leader behavior for the *i*th CEO. Greater linear pattern fit is indicated by larger positive $b_{y_i.x_i}$. We used the unstandardized slopes, as opposed to the standardized slopes, because the unstandardized slopes are unaffected by differential variances in the ratings of the 21 primary leadership dimensions across CEOs.

In our Table 5.1 example, the pattern similarity fit is measured by the within CEO linear regression slopes for each CEO. These calculated slopes indicate that CEO A has an unstandardized regression weight of 1.29, CEO B has a weight of 1.14, and the weight for CEO C is .50. The larger the *b* weight, the greater the pattern similarity fit between the CLT and CEO behavior. Therefore, the closest pattern fit occurred for CEO A, CEO C had the worst fit, and CEO B, is pretty good as this CEO had an in-between level of fit but closer to A than C.

Absolute Behavioral Fit

The second aspect of fit was an assessment of absolute level of agreement between CLTs and behavior. This aspect of fit was conceptualized and measured as the square root of the absolute reliability coefficient (ρ_i) discussed in generalizability theory. The absolute reliability coefficient indicates the ability of the CLT ratings to exactly predict the level of the CEO's leadership behaviors.

The absolute reliability coefficient was computed by conducting a completely randomized factorial ANOVA for each CEO. In this factorial ANOVA, one facet was the attribute of leadership being rated (i.e., 21 primary leadership dimensions) and the other facet was "leadership CLTs versus leader behavior." The dependent variable for this analysis was a newly created variable we labeled "rating." This variable consisted of 42 lines of data for each CEO with the first 21 being the House and colleagues' (2004) country-level 21 CLT dimension ratings and the second 21 lines being the 21 average behavior ratings for that CEO. A source table was calculated for each CEO and the variance components of the design were computed. This analysis enabled us to calculate the absolute reliability coefficient (ρ_i) as follows:

$$\rho_i = \frac{\sigma^2_{Leader\,Profile}}{\sigma^2_{Leader\,Profile} + \sigma^2_{CLT} + \sigma^2_{Leader\,Profile*CLT}} \tag{5.5}$$

Where $\sigma^2_{Leader\,Profile}$ is the variance estimate for the effect of the leadership profile; σ^2_{CLT} represents the variance for the CLT prediction, and $\sigma^2_{Leader\,Profile*CLT}$ is the interaction between the leadership profile and the CLTs. The higher the ρ_i, the better the agreement between CLTs and CEO behavior. Mathematically, ρ_i is a variance estimate and so we took the square root of this estimate to yield the reliability index. This reliability index is basically a correlation, but it is important to remember that it expresses the relationship between the CLT to perfectly capture the CEO behavior.

Returning to our simulated example in Table 5.1, our calculation for the absolute behavioral fit index showed the closest square root absolute fit occurred for CEO C (.92), followed by CEO B (.67). The absolute behavior fit index was .39 for CEO A. Note, the rank order of fit according to this absolute behavior fit agreement is C, followed by B then A. This ranking is in contrast to the rank order of fit as provided by the pattern fit index (i.e., A was the highest, followed by B, then C).

To obtain a single measure of fit, we combined these two pieces of information into a single composite measure of fit. We created this single composite fit measure by first standardizing the pattern fit index and the square root of the absolute reliability coefficient and then averaging these two standardized indices into a single index. This standardization process allows us to create a composite measure of fit that equally weights both aspects of fit. The rank order for this combined index for the simulated data (assuming a mean and standard deviation [*SD*] of this simulated data of .6 and .4, respectively, for both aspects of fit) are .60, .76, and .28. On the basis of these scores, the best order of overall fit is CEO B, CEO A, and then CEO C. The reason CEO B had the highest fit was that this CEO was in the middle position for pattern fit and absolute behavioral fit. On the other hand, CEOs A and C had substantially poorer fit with at least one aspect of the Gestalt Fit measure thereby pulling their position on the overall fit score down. Further examples, explanation, and illustrations of pattern, behavioral similarity, and Gestalt Fit will be provided in Chapter 9.

In summary, in this chapter we discussed the rationale and procedure for developing scales and testing our hypotheses. We accomplished the following:

- Indicated which statistical methods would be used to test the viability of our conceptual models as well as the methods that would be used to test our hypotheses
- Discussed the necessity of justifying aggregation to the CEO level of analysis
- Indicated the various psychometric analyses that would be performed and included a discussion of measurement equivalence issues and analysis that would be performed to develop our measures
- Explained what HLM statistical analysis is and why we used it to test our hypotheses.
- Described our new measure of Gestalt Fit, which we used in this study to test our theoretical proposition that the match between country-level leadership expectations (CLT) and actual CEO behavior leads to critical outcomes
- Presented an example of Gestalt Fit, providing an illustration of how the two constituent components combine into a single overall fit index