CHAPTER 7

EVALUATION OF MODEL FIT AND ADEQUACY

D. Betsy McCoach and Anne C. Black

"All models are false, but some are useful."

—Box, 1979, p. 202

How do researchers evaluate multilevel models? How should they choose among competing models? The utility of any model depends upon its ability to explain the phenomenon under investigation. Therefore, assessment of model adequacy should consider two aspects of the model: 1) model fit, or the use of model selection criteria to choose among competing models, and 2) the explanatory power of the model, or the ability of the predictors to explain scores on the outcome variable. While model fit is evaluated relative to other competing models, the explanatory power of the model may be evaluated both relative to competing models and in an absolute sense (i.e., does the model do a good or poor job of explaining scores on the outcome variable?). This chapter explains common measures of model adequacy within the multilevel modeling literature. Further, we briefly describe several areas of controversy or confusion surrounding measures of model adequacy. Finally, we provide recommendations for evaluating the adequacy of a multilevel model.

Multilevel Modeling of Educational Data, pages 245–272

MODEL SELECTION CRITERIA

Model selection should be guided by theory and informed by data. Burnham and Anderson (2004) suggest that three general principles should guide model selection in the social sciences. First, parsimony is paramount. Adding additional parameters is likely to improve fit and cannot lead to worse model fit (Forster, 2000). The critical issue is whether the improvement in the fit of the model justifies the inclusion of the additional parameters. Second, Burnham and Anderson advocate the use of multiple working hypotheses. Using data to compare several plausible competing hypotheses often provides more useful information than comparing a given model to an often implausible null hypothesis. Third, one of the central tenets of scientific research is the use of quantitative information to judge the strength of evidence (Burnham & Anderson). Finally, researchers should examine the model to ensure that the estimated parameters make sense and seem plausible.

Model selection is a crucial part of the multilevel modeling process. How does the researcher select the appropriate model from among several competing models? Model selection requires striking a delicate balance between parsimony and complexity. The researcher's goal is to "arrive at a model that describes the observed data to a satisfactory extent but without unnecessary complications" (Snijders & Bosker, 1999, p. 91). The most common methods of model selection include hypothesis testing approaches and "information criteria," or index comparison, approaches. After briefly reviewing the concept of deviance, we explain how to use the chi-square difference test to compare the deviances of two nested models. We then review the use of index comparison approaches, such as the AIC and BIC, for model selection. Finally, we explain the use of R-squared type measures to determine the predictive power of the multilevel model. Researchers must consider both the fit and the predictive ability of a given multilevel model to determine the adequacy of the model.

Deviance

Maximum likelihood estimation techniques provide estimates for the values of the population parameters that maximize the probability of obtaining the observed data (Singer & Willett, 2003). A likelihood function "describes the probability of observing the sample data as a function of the model's unknown parameters" (Singer & Willett, p. 66). The parameter estimates are those estimates that maximize the likelihood function. When we use maximum likelihood (ML) to estimate the parameters of the model, the estimation also provides the likelihood, which easily can be transformed into a deviance statistic (Snijders & Bosker, 1999).

The deviance compares the log-likelihood of the specified model to the log-likelihood of a saturated model that fits the sample data perfectly (Singer & Willett, 2003, p. 117). Specifically, deviance = –2 (log-likelihood of the current model—log-likelihood of the saturated model) (–2LL) (Singer & Willett). Therefore, deviance is a measure of the badness of fit of a given model; it describes how much worse the specified model is than the best possible model (Singer & Willett). Deviance statistics cannot be interpreted directly since deviance is a function of sample size as well as the fit of the model. However, differences in deviance can be interpreted for competing models, if those models are hierarchically nested, use the same data set, and use full maximum likelihood estimation techniques to estimate the parameters.

In full maximum-likelihood estimation, the estimates of the variance and covariance components are conditional upon the point estimates of the fixed effects (Raudenbush & Bryk, 2002). When using full maximum likelihood (FIML), the number of parameters includes both the fixed effects and the variance and covariance components. Restricted maximum likelihood (REML) estimates of variance-covariance components adjust for the uncertainty about the fixed effects; FIML estimates do not (Raudenbush & Bryk). When the number of level-two units is large, REML and FIML results will produce similar estimates of the variance components. However, when there are few level-two units, the maximum likelihood estimates of the variance components (τ_{qq}) will be smaller than those produced by REML, and the REML results may be more realistic (Raudenbush & Bryk). The deviances of any two nested models that differ in terms of their fixed and/or random effects can be compared when using FIML. However, REML only allows for comparison of nested models that differ in their random effects (Snijders & Bosker, 1999, p. 89).

Hypothesis Testing

Hypothesis testing is one of the most commonly utilized model selection methods (Weaklim, 2004). In multilevel modeling, researchers often use chi-square difference tests to compare the fit of two different models. In addition, hypothesis tests are used to evaluate whether fixed effects, random level-one coefficients, and variance components are statistically significantly different from zero (Raudenbush & Bryk, 2002). Finally, general linear hypothesis testing using the Wald statistic allows researchers to test composite hypotheses about sets of fixed effects (Singer & Willett, 2003). Because this chapter is devoted to the determination of model fit issues, we focus our attention on the use of chi-square difference tests to determine the adequacy of the multilevel model.

Chi-square Difference Test

Two models are nested when one model is a subset of the other (Kline, 1998). In other words, in nested models, "the more complex model includes all of the parameters of the simpler model plus one or more additional parameters" (Raudenbush, Bryk, Cheong, & Congdon, 2000, p. 80–81). If two models are nested, the deviance statistics of two models can be compared directly. The deviance of the simpler model (D_1) minus the deviance of the more complex model (D_2) provides the change in deviance ($\Delta D = D_1 - D_2$). The simpler model always will have at least as high a deviance as the more complex model, and generally the deviance of the more complex model will be lower than that of the simpler model. In large samples, the difference between the deviances of two hierarchically nested models is distributed as an approximate chi-square distribution with degrees of freedom equal to the difference in the number of parameters being estimated between the two models (de Leeuw, 2004). We refer to the number of parameters in the larger (less parsimonious) model as p_1 and the number of estimated parameters in the smaller (more parsimonious) model as p_2.

In evaluating model fit using the chi-square difference test, the more parsimonious model is preferred, as long as it does not result in significantly worse fit. In other words, if the model with the larger number of parameters fails to reduce the deviance by a substantial amount, the more parsimonious model is retained. Therefore, when the change in deviance (ΔD) exceeds the critical value of chi-square with ($p_1 - p_2$) degrees of freedom, the difference in the deviances is statistically significant. In this situation, we favor the more complex model. However, if the more complex model does not result in a statistically significant reduction in the deviance statistic, we favor the more parsimonious model.

Full maximum likelihood estimation maximizes the likelihood of the sample data, whereas restricted maximum likelihood estimation maximizes the likelihood of the residuals (Singer & Willett, 2003, p. 118). In FIML, the number of reported parameters includes the fixed effects (the γ's) as well as the variance/covariance components. When using REML, the number of reported parameters includes only the variance and covariance components. To compare two nested models that differ in their fixed effects, it is necessary to use FIML estimation, not REML estimation. REML only allows for comparison of models that differ in terms of their random effects but have the same fixed effects. Because most programs use REML as the default method of estimation, it is important to remember to select FIML estimation to use ΔD to compare two hierarchically-nested models with different fixed effects.

An Example

Consider the following model:

$$\begin{aligned}Y_{ij} &= \beta_{0j} + \beta_{1j}(IQ)_{ij} + r_{ij} \qquad (7.1)\\ \beta_{0j} &= \gamma_{00} + \gamma_{01}(SchoolSES)_j + u_{0j}\\ \beta_{1j} &= \gamma_{10} + \gamma_{11}(SchoolSES)_j + u_{1j}\end{aligned}$$

Remember, the number of estimated parameters in FIML is equal to the number of fixed effects (γ's) plus the number of variance covariance components. In this example, there are four fixed effects (γ_{00}, $\gamma_{01,}$ $\gamma_{10,}$ and γ_{11}). In addition, there are four variance covariance components (σ^2, the variance of r_{ij}; τ_{00}, the variance of u_{0j}; τ_{11}, the variance of u_{1j}; and τ_{01}, the covariance of u_{0j} and u_{1j}.). Therefore, there are eight estimated parameters in FIML. In contrast, the number of estimated parameters in REML is simply the number of variance/covariance components (σ^2, the variance of r_{ij}; τ_{00}, the variance of u_{0j}; τ_{11}, the variance of u_{1j}; and τ_{01}, the covariance of u_{0j} and u_{1j}.). In this example, there are four estimated parameters in REML.

Imagine we wanted to compare the model above to the following model, a model in which the SES/IQ slope remains constant across schools:

$$\begin{aligned}Y_{ij} &= \beta_{0j} + \beta_{1j}(IQ)_{ij} + r_{ij} \qquad (7.2)\\ \beta_{0j} &= \gamma_{00} + \gamma_{01}(SchoolSES)_j + u_{0j}\\ \beta_{1j} &= \gamma_{10} + \gamma_{11}(SchoolSES)_j\end{aligned}$$

We are no longer estimating a variance for u_{1j} or the covariance of u_{1j} and u_{0j}. Therefore, model 2 contains six estimated parameters in FIML and two estimated parameters in REML. The difference between the deviance of model 1 and model 2 could be compared using either REML or FIML since the two models vary only in their variance-covariance components. Assume that the deviance of model 1 is 32, and the deviance of model 2 is 45; therefore, the difference between the deviances is 13. We compare this to the critical value of χ^2 with two degrees of freedom (which is 5.99). Because 13 is larger than 5.99, we reject the null hypothesis that the simpler model provides an equally good fit to the data, we determine that the simpler model fits significantly worse than the more complex model (the more complex model fits significantly better than the simpler model). Therefore, we conclude that we cannot make the proposed simplifications, and we opt in favor of the more complex model.

Finally, consider model 3, as compared to our initial model, model 1.

$$\begin{aligned} Y_{ij} &= \beta_{0j} + \beta_{1j}(IQ)_{ij} + r_{ij} \qquad (7.3) \\ \beta_{0j} &= \gamma_{00} + \gamma_{01}(SchoolSES)_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \end{aligned}$$

Model 3 eliminates the cross-level interaction between School SES and IQ. Model 3 contains seven estimated parameters in FIML (three fixed effects: γ_{00}, $\gamma_{01,}$ and $\gamma_{10,;}$ and four random effects: σ^2, τ_{00}, τ_{11}, and τ_{01}.) However, model 3 has four estimated parameters in REML, just as model 1 did. This demonstrates that models 1 and 3 are nested models in FIML but not in REML.

Other Model Selection Techniques

Using hypothesis testing procedures is one of the most commonly employed model selection methods in multilevel modeling. However, the hypothesis testing approach to model selection has been criticized on several grounds (Raftery, 1995). First, with large sample sizes, most null hypotheses are rejected. Therefore, the use of hypothesis tests for model selection can produce very complex models (Weaklim, 2004). Second, when a given model is selected from multiple models, *p*-values do not have the same interpretation as they do when only two models are considered, and *p*-values can be misleading in this situation (Raftery). "By choosing among a large number of variables, one increases the probability of finding 'significant' variables by chance alone" (Raftery, p. 118). In addition, classical hypothesis tests do not necessarily identify a single best model (Weaklim). While significance tests permit the researcher to reject or fail to reject the null hypothesis, significance tests do not actually provide evidence in support of the null hypothesis (Raftery). In other words, we have no evidence that the model we failed to reject is better than or preferable to the comparison model; we only can say that it is not significantly worse than the comparison model. Therefore, the more parsimonious model can be rejected, but it can never be 'confirmed' (Weaklim). By convention, researchers generally choose the most parsimonious model that is not rejected, selecting the simplest model unless there is statistical evidence suggesting the more complex model is preferable. However, this process cannot answer the question of which model is better. "Null hypothesis testing only provides arbitrary dichotomies (e.g., significant vs. non-significant), and in the all-too-often-seen case in which the null hypothesis is false on a priori grounds, the test result is superfluous" (Burnham & Anderson, 2004, p. 266). Since all models are simplifications of reality, all models are likely to be misspecified to a certain degree. Therefore, hypothesis tests do not provide guidance to help select an *imperfect* but parsimonious model (Gelman & Rubin, 1995). Further, hypothesis tests do not aid researchers in deciding whether the lack of fit of a parsimonious model is a problem in practice (Gelman & Rubin). Finally, hypothesis testing procedures do not quantify how *much* better a particular model is. Because hypothesis testing does not allow us to quantify the degree of fit or misfit; we cannot quantify the degree to which one model should be preferred over another.

However, the largest drawback of the hypothesis testing approach is that it only permits the comparison of nested models. It is often impossible to compare competing hypotheses using nested statistical models (Raftery, 1995). This is especially true when the models embody dissimilar or contradictory views of the process or phenomenon under examination (Raftery). Because hypothesis testing procedures only allow for comparison of nested models, if we wish to compare two models with different sets of predictors, we cannot use the chi-square difference test or any other hypothesis testing procedure. In this situation, model selection indices, such as the Akaike Information Criterion (*AIC*) and the Bayesian Information Criterion (*BIC*), are particularly helpful because they allow us to rank or compare models with different sets of parameters.

AIC and BIC

While model index comparison approaches, such as *AIC* and *BIC*, have received relatively little attention within the educational literature, their use is quite common within the sociological literature. (See, for example, *Sociological Methods and Research*, Volume 33(2), 2004, a special issue devoted to model selection issues in sociology). Information theoretic model selection represents, in some sense, the converse of classical hypothesis testing procedures (Bozdogan, 1987). Information theoretic techniques focus on "choosing a critical value which then determines, approximately, what the significance level is or might be" (Bozdogan, p. 363); whereas, in statistical significance testing, the researcher sets the probability of Type I error (alpha), which then determines the critical value.

There are several advantages to using the *AIC* or the *BIC* rather than relying upon deviance statistics and chi-square difference tests to evaluate the goodness of fit of a multilevel model. First, the *AIC* and *BIC* allow the comparison of non-nested models. As long as the sample remains constant, *AIC* and *BIC* allow the comparison of competing models, whether or not they are hierarchically nested. Further, selection indices such as the *AIC* and the *BIC* quantify the degree to which the given model represents an improvement over comparison models.

The Bayesian approach to model selection regards every competing model as the possible "true" model, and then estimates the likelihood that the model in question is, indeed, the correct model (Zucchini, 2000). For the *AIC*, the prediction of future data is the key criterion of the adequacy of a model (Kuha, 2004). Therefore, although the formulas for the *BIC* and the *AIC* appear similar, the philosophical underpinnings of the two approaches differ dramatically. The field of sociology tends to favor the *BIC*; whereas, econometricians tend to prefer the *AIC* (Kuha, 2004). Whether researchers should use the *AIC* or the *BIC* for model selection purposes has been the subject of much debate and scrutiny (Kuha, 2004; Weaklim, 1999, 2004). We choose to sidestep the controversy surrounding the choice of the *AIC* or the *BIC*. Rather, we believe that the combined use of the *AIC* and the *BIC* (in conjunction with chi-square difference tests for nested models) can be quite informative. While our explanations of the *AIC* and *BIC* are conceptually and mathematically shallow, we believe that they will serve the applied researcher. Those interested in the conceptual and methodological underpinnings of the *AIC* and the *BIC* should refer to Bozdogan (1987), Burnham and Anderson, (2004), Raftery (1995), Schwarz (1978), Wagenmeyers and Farrell (2004), and Zucchini (2000).

Not all software programs provide *AIC* and *BIC* measures in their output. HLM 6.04 does not provide estimates of *AIC* and *BIC*; however, SPSS, SAS, R, and MPLUS do provide these indices. Both the *AIC* and the *BIC* can be computed easily from the deviance statistic. Because *AIC* and *BIC* are computed from the deviance statistic, FIML generally is considered the most appropriate estimation method to use when computing information criteria (Verbeke & Molenberghs, 2000). However, a recent simulation study by Gurka (2006) suggests that information criteria such as the AIC and BIC also may perform at least as well under REML as they do under FIML. Further research is needed to determine whether information indices such as the AIC and BIC can be used with REML, but Gurka's results suggest that the use of information criteria under REML may not be as problematic as was once believed.

The Akaike Information Criterion (AIC). The formula for the *AIC* is shown below.

$$AIC = D + 2p \tag{7.4}$$

where D is deviance and p = the number of parameters estimated in the model.

To compute the *AIC*, simply multiply the number of parameters by two and add this product to the deviance statistic. As you will recall, the deviance (or –2log-likelihood [–2LL]) represents the degree "of inaccuracy, badness of fit, or bias when the maximum likelihood estimators of the parameters of a model are used" (Bozdogan, 1987, p. 356). The second term, $2p$, imposes a penalty based on the complexity of the model. When there are several competing models, the model with the lowest *AIC* value is considered to be the best model. Because the *AIC*'s penalty term is equal to $2p$, the deviance must decrease by more than 2 per additional parameter in order to favor the model with greater numbers of parameters.

Compare this to the chi-square difference test for model selection. The critical value of χ^2 with one degree of freedom at $\alpha = .05$ is 3.84. Therefore, when comparing two models that differ by one degree of freedom, the chi-square difference test actually imposes a more stringent criterion for rejecting the simpler model. In fact, this is true for comparisons of models that differ by seven or fewer parameters. Therefore, using the chi-square difference test will result in an equivalent or more parsimonious model than using the *AIC* when comparing models that differ by seven or fewer parameters. However, when comparing models that differ by more than seven parameters, the *AIC* will favor more parsimonious models.

The Bayesian Information Criterion (BIC). The *BIC* is equal to the sum of the deviance and the product of the natural log of the sample size and the number of parameters. The formula for the *BIC* is shown below.

$$BIC = D + \ln(n) * p \tag{7.5}$$

where D is deviance (–2LL),
p = the number of parameters estimated in the model, and
n = the sample size.

Therefore, the *BIC* imposes a penalty on the number of parameters that is impacted directly by the sample size. In multilevel models, it is not entirely clear which sample size should be used with the BIC: the number of units at the lowest level, the number of units at the highest level, or some weighted average of the two. SAS PROC MIXED uses the number of independent sampling units as the sample size when computing the *BIC*. In contrast, SPSS and R use the level-one sample size in their computation of the *BIC*. Therefore, even though SPSS and SAS will produce identical –2LL and AIC values, the *BIC* value will differ across these programs. Since the *BIC* imposes a steeper per parameter penalty as the sample size increases, the *BIC* value produced by SPSS and R will be larger than the *BIC* value produced by SAS, and it will tend to favor more parsimonious models. MPLUS also uses the level-one sample size in the computation of *BIC*. However, because growth models in MPLUS would typically be formulated in the wide or multivariate format, the effective sample size used for the computation of the

BIC in MPLUS is the number of people in the sample. Thus, the choice of sample size for the computation of the *BIC* is not without controversy. Future research should address the impact of this choice on model selection. In the meantime, researchers should carefully consider which sample size they are implicitly or explicitly using in their computation of the *BIC*.

However, even at small sample sizes, the *BIC* will favor more parsimonious models than the *AIC* or traditional chi-square difference tests. Given a sample size as low as 50, the penalty for the *BIC* is 3.912 times the number of parameters. In contrast, the penalty for the *AIC* is two times the number of parameters, and the rejection region for traditional chi-square difference tests is 3.84 for one parameter, 5.99 for two parameters, etc.

The model with the lowest *BIC* is considered to be the best fitting model. Raftery (1995) provided guidelines for interpreting changes in *BIC*. Subtract the *BIC* for model 2 from the *BIC* for model 1 to compute a *BIC* difference ($BIC_1 - BIC_2$). Raftery suggests that *BIC* differences of 0–2 provide weak evidence favoring model 2 over model 1, *BIC* differences of 2–6 provide positive evidence for favoring model 2, *BIC* differences of 6–10 provide strong evidence favoring model 2, and *BIC* differences above 10 provide very strong evidence favoring model 2 over model 1.

The central question in model selection is how much additional information a given parameter must add to justify its inclusion (Weaklim, 2004). Figure 7.1 provides a graph of the decision values for the chi-square differ-

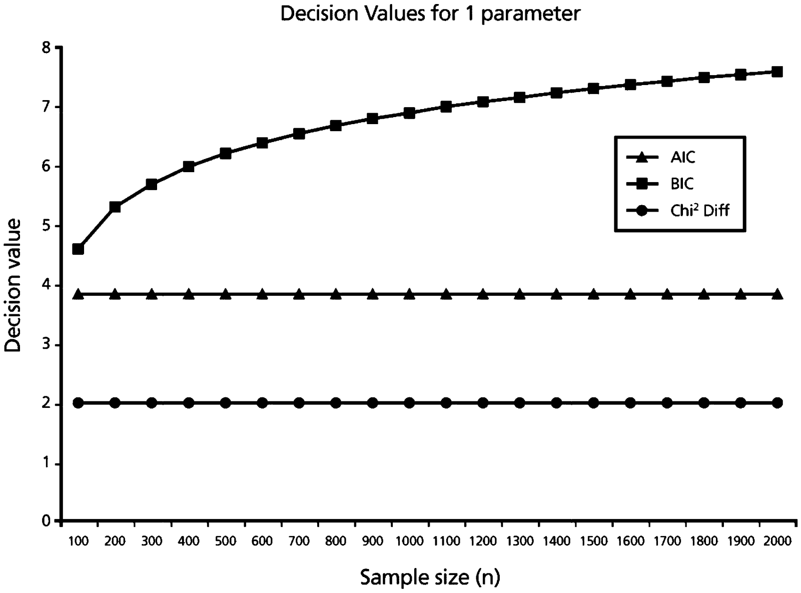

Figure 7.1 Critical decision values for *AIC*, *BIC*, and chi-square difference measures with successive increases in sample size of 100.

ence test, the *AIC*, and the *BIC* for two nested models that differ by one parameter at a variety of sample sizes. The y-axis shows the change in deviance necessary to favor the more complex model over the simpler model. In the figure, this quantity is called the "decision value." The figure clearly shows that for one-parameter tests, the *AIC* placed the least stringent criterion for favoring the more complex model. The *BIC* places the most stringent criterion for favoring the more parsimonious model, and the penalty that the *BIC* imposes becomes increasingly stringent as the sample size increases. However, the relationship between the *BIC* and the sample size is curvilinear, and increasing the sample size has very little impact on the decision value of the *BIC* once the sample reaches 10,000 or greater. For example, the penalty for a sample size of 10,000 is 9.21; the penalty for a sample size of 20,000 is 9.90; the penalty for a sample size of 30,000 is 10.31.

Figure 7.2 provides a graph of the decision values for the chi-square difference test, the *AIC*, and the *BIC* with sample sizes of 200, 1000, and 5000 for nested models as a function of the change in the number of parameters. When comparing models that differ by a small number of parameters (seven or fewer), the *AIC* will result in the most complex models. The *BIC* will always favor more parsimonious models than the *AIC* or the chi-square difference test, and this effect is especially pronounced at larger sample sizes.

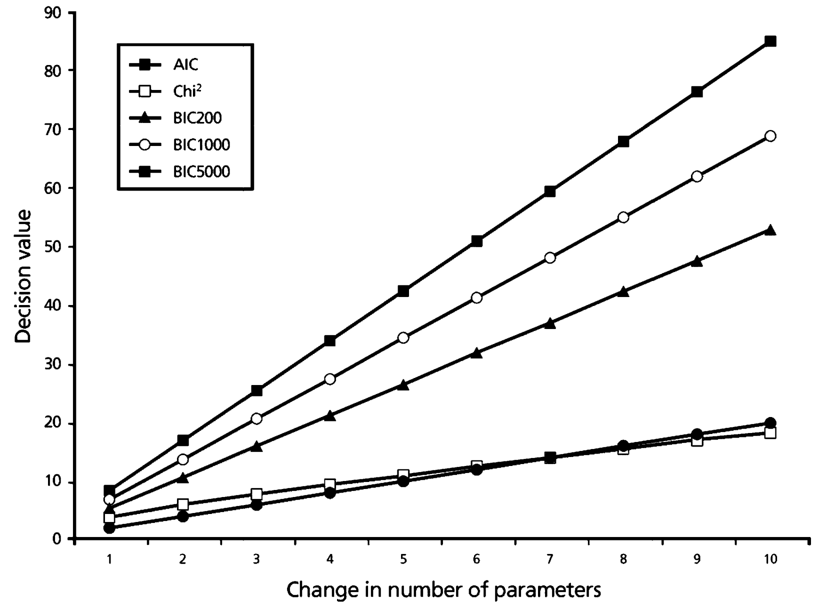

Figure 7.2 Critical decision values for *AIC*, *BIC*, and chi-square difference measures where $n = 200$, 1000, and 5000, as number of parameters increases from 1 to 10.

In conclusion, of the three model selection techniques, *BIC* favors the most parsimonious models, regardless of sample size. The *AIC* and the chi-square difference test often will provide similar results. However, the chi-square difference test favors more parsimonious models for smaller changes in the number of parameters, and the *AIC* favors more parsimonious models in the case of larger changes in the number of parameters. The *BIC* explicitly takes sample size into consideration, while chi-square and *AIC* do not. It is important to remember that the chi-square difference test only can be used to compare nested models. However, the *AIC* and the *BIC* can be used to compare both nested and non-nested models. Our recommendation is to examine *AIC*, *BIC*, and chi-square difference tests (for nested models). Most of the time, the three methods will converge upon the same decision. When there is a discrepancy among the three indices, we recommend using professional judgment and knowledge of the research area to guide decision making.

Example

To illustrate the use of the hypothesis testing and model selection approaches, we turn our attention to Table 7.1, which contains fixed and random effects estimates for six different models, and to Table 7.2, which contains their associated fit statistics. We illustrate a model-building approach to model the between- and within-school variability of students' reading achievement at the beginning of kindergarten using the Early Childhood Longitudinal Study, Kindergarten Cohort (ECLS-K). (See chapters 2 and 6 of this volume (Stapleton & Thomas, 2008, and O'Connell, Goldstein, Rogers, & Pens, 2008, respectively) for details about the ECLS-K dataset.) The sample size for this analysis includes 7215 first-time kindergarteners in 578 kindergartens. To enable comparison of models that differed in their fixed effects using the chi-square difference test and to compute the *BIC* and *AIC* values, we estimated all models using FIML. The models in Table 7.1 utilize a small set of student- and school-level predictors to explain students' reading achievement in the fall of the kindergarten year. Student-level variables include age and SES. School-level variables include school type (public/private), the percentage of students who receive free lunch in the school (a measure of school *SES*), and the percentage of minority students in the school. Model 1 is the baseline model. As a control variable, it includes the number of months of kindergarten the student attended prior to taking the reading achievement test. Model 2 adds the student's age at kindergarten entry as a level-one predictor. As seen in Table 7.1, the fixed effect of SES is statistically significant (the parameter estimate is .36; the standard error is

TABLE 7.1 Fixed Effects and Variance Components for ECLS-K Example Analyses

Fixed effects	Parameter	Model 1 (baseline)	Model 2 (adds age of kindergarten entry)	Model 3 (adds SES)	Model 4 (SES—with no random slope)	Model 5 (adds sector and % free lunch)	Model 6 (adds % minority)
For Intercept (β_0)							
Intercept	γ_{00}	20.86 (.73)	20.88 (.72)	19.97 (.55)	20.09 (.56)	19.22 (.52)	19.23 (.52)
Private	γ_{01}					1.08 (.35)	1.06 (.35)
% free lunch	γ_{02}					–.05 (.01)	–.05 (.01)
% minority	γ_{03}						.003 (.01)
For SES slope (β_1)							
Intercept	γ_{10}			3.91 (.15)	3.88 (.13)	3.71 (.18)	3.71 (.18)
Private	γ_{11}					–1.38 (.40)	–1.4 (.41)
% free lunch	γ_{12}					–.02 (.01)	–.025 (.01)
% minority	γ_{13}						.003 (.01)
For K exp. (β_2)							
Intercept	γ_{20}	1.28 (.33)	1.28 (.33)	1.73 (.25)	1.74 (.25)	1.92 (.23)	1.92 (.23)
For entry age (β_3)							
Intercept	γ_{30}		.36 (.02)	.36 (.02)	.36 (.02)	.35 (.02)	.36 (.02)
Public	γ_{31}						
Var. comp.							
	τ_{00}	13.79 (1.13)	13.52 (1.11)	5.19 (.61)	5.38 (.59)	3.80 (.52)	3.78 (.52)
	τ_{01}			2.19 (.45)		2.01 (.41)	1.99 (.41)
	τ_{11}			2.19 (.68)		2.17 (.65)	2.15 (.65)
	σ^2	60.48 (1.05)	58.70 (1.01)	54.61 (.97)	55.47 (.96)	54.32 (.96)	54.33 (.96)
Number of parameters		4	5	8	6	12	14

TABLE 7.2 Pseudo-R^2 and Model Fit Measures for the ECLS-K Examples

Pseudo-R^2 and model fit	Model 1 (baseline)	Model 2 (adds age of kindergarten entry)	Model 3 (adds SES)	Model 4 (SES—with no random slope)	Model 5 (adds sector and % free lunch)	Model 6 (adds % minority)
Proportional reduction in variance—Level one		.03	.10	.08	.10	.10
Prop. reduction in prediction error—Level one		0.03	.19	0.18	0.22	0.22
Proportional reduction in variance—Level two		0.02	.62	0.61	0.72	0.73
Prop. reduction in prediction error—Level two		0.02	.48	0.47	0.56	0.56
Level-two R^2 slope (Raudenbush & Bryk, 2002)					.009	.02
Deviance	50,819.43	50,608.21	49,847.04	49,890.46	49,725.25	49,724.82
AIC	50,827.43	50,618.21	49,863.04	49,902.46	49,749.25	49,752.82
BIC (n= the number of level-2 units)	50,844.87	50,640.01	49,897.92	49,928.62	49,801.57	49,813.86
BIC (n= the number of level-1 units)	50,854.966	50,662.63	49,918.11	49,943.76	49,831.86	49,848.58
Number of parameters	4	5	8	6	12	14

.02). However, we are most interested in using the change in deviance and the *AIC* and *BIC* to assess the fit of the two models.

To use the chi-square difference test to compare the two models, we compute the change in deviance (ΔD) and the change in the number of estimated parameters (Δp) and compare these values to the critical value of chi-square with Δp degrees of freedom. First, we compare model 1, the baseline model, to model 2, a model which adds (grand-mean centered) age of kindergarten entry as a level-one predictor. The change in deviance is 50,819.43 – 50,608.21 = 211.22. The change in the number of parameters is 5 – 4 = 1. We compare 211.22 to the critical value of chi-square with one degree of freedom, which is 3.84. Because 211.22 > 3.84, we reject the null hypothesis that the two models fit the data equally well, and we favor the more complex model. The additional parameter results in improved model fit.

Using *AIC*, we favor the model with the smaller *AIC* value. Table 7.2 shows that the deviance for model 1 is 50,819.43 with 4 parameters. Therefore the *AIC* is 50,819.43 + 2*4, or 50,827.43. The *AIC* for model 2 is 50,608.21 + 5*2, or 50,618.21. Using *AIC*, we conclude that the model that includes age is superior to the model that does not include age.

Finally, we compare the *BIC* values for the two models. For these examples, we used the number of level-two units as the sample size for the computation of the *BIC*. For completeness and for comparison purposes, Table 7.2 provides 2 different *BIC* estimates: the *BIC* computed using the number of level-2 units as the effective sample size and the *BIC* computed using the number of level-1 units as the effective sample size.

Using n = the number of level 2 units, the *BIC* for model 1 in Table 7.2 is 50,819.43 + 4*ln(578), or 50,844.87. The BIC for model 2 is 50,608.21 + 5*ln(578), or 50,640.01. The *BIC* for model 2 is smaller than the *BIC* for model 1, so we again conclude that model 2 provides better fit to the data than model 1. In addition, the change in *BIC* (204.86) is greater than 10. Therefore, according to Raftery's (1995) rules of thumb, the difference in *BIC* provides very strong evidence for favoring model 2 over model 1. In this case, we would draw the same conclusions regarding model selection if we were to use the *BIC* computed using the level-one sample size.

Model 3 includes SES as a student-level predictor of beginning kindergarten reading. The slope of SES is random; therefore, the inclusion of SES adds three additional parameters to the model: one fixed effect (γ_{10}) and two random effects (τ_{01} and τ_{11}). The fixed effect of SES is statistically significant, and both additional random effects (τ_{01} and τ_{11}) are also statistically significant. However, to assess the model fit, we examine the effects of adding these three parameters on the change in deviance, the *AIC*, and the *BIC*. Adding these three additional parameters decreases the deviance from 50,608.21 to 49,847.04. This change in deviance of 761.17 exceeds the critical value of 7.82, the critical value of chi-square with three degrees of freedom at $\alpha = .05$. In addition, the *AIC* is substantially smaller in the model that includes SES (49,863.04) than in the model without SES (50,618.21). Finally, the change in *BIC* from model 2 (50,640.01) to model 3 (49,897.92) is 742.09. This difference in *BIC* provides very strong evidence for favoring model 3 over model 2 (Raftery, 1995).

Finally, let us examine what happens when we add the percentage of minority students as a level-two (school-level) predictor, and compare the resulting model (model 6) to the previous model (model 5). First, the percentage of minority students does not exert a statistically significant influence on either the intercept (γ_{03}) or on the SES slope (γ_{13}). When we compare the two models using the chi-square difference test, the change in deviance is .42 (49,725.25 – 49.724.82) for a two-parameter change in the model. This is below the critical value of chi-square with two degrees of freedom (5.99); therefore, we fail to reject the null hypothesis and conclude that the more parsimonious model (without percentage minority students) does not provide a statistically significantly poorer fit to the data than the model that includes those two parameters. Further, the *AIC* is smaller for the model that does not include percentage of minority students (49,749.25 for model 5 vs. 49,752.82 for model 6). Finally, the *BIC* is smaller for model 5 (49,801.57) than for model 6 (49,813.86), and this difference (12.29) is larger than 10. Therefore, this difference in *BIC* provides very strong evidence for favoring model 5 over model 6 (Raftery, 1995).

Summary—Model Selection Criteria

These examples, in combination with Figures 7.1 and 7.2, demonstrate that much of the time, the chi-square difference test, the *AIC*, and the *BIC* will converge, and point toward the selection of the same model. In other situations, the *AIC*, *BIC*, and chi-square difference tests may lead to conflicting conclusions. When these results diverge, the researcher must make a difficult decision about whether he or she favors model parsimony or model complexity. In borderline cases in which the change in deviance between two models is relatively small, the *BIC* favors the more parsimonious model while the *AIC* (and the chi-square difference test) favors the more complex model. In these situations, it is very important for the researchers to use their substantive knowledge and judgment to reach a conclusion about the "best model." Kuha (2004) suggests that when the *AIC* and the *BIC* err, "the *AIC* tends to favor models that are too large, and *BIC* models that are too small. Thus, an optimistic interpretation of these results is that even a disagreement at least suggests bounds for the range of acceptable models" (Kuha, p. 222). Very little research has specifically examined the AIC and BIC within a multilevel framework. However, Whittaker & Furlow (2006) conducted a simulation study to examine the performance of the AIC and the BIC under a number of different conditions when estimating two-level hierarchical linear models. They found that when the information indices did not select the correct model, the AIC tended to select the more parameterized (less parsimonious) model, whereas the BIC tended to select the less parameterized (more parsimonious) model. In this situation, we suggest considering substantive and theoretical issues as well as empirical and statistical ones. Cudeck and Henley (1991) also provide advice that is useful in this regard. They suggest that when evaluating the relative performance of competing models, often the best that can be done is to state clearly the criteria that are used in the comparison, in conjunction with descriptions of the models, characteristics of the data, and the purpose for which the models were constructed; moreover, this is actually a useful accomplishment whose value should not be minimized. Finally, when two competing models appear to fit the data (almost) equally well, replication studies using a new sample may be the most effective way to determine which model is truly "the best."

In conclusion, it is important to remember that the researcher plays an important role in the evaluation of model fit. No mechanical data analytic procedure for evaluating model fit should override human judgment (Browne, 2000). Examining changes in deviance in combination with model selection indices, such as the *AIC* and the *BIC*, seems to be the most prudent course of action for evaluating model fit. The rules of thumb presented earlier provide guidance for the researcher; however, they should not be used blindly or mechanically.

REDUCTION IN VARIANCE ESTIMATES

Complementary to the use of model fit criteria for model evaluation and selection, analogs of the squared multiple correlation, R^2, may be used to assess the ability of a given model to explain the data. In ordinary multiple regression, the squared multiple correlation between the outcome variable and the linear combination of weighted predictor variables represents the proportion of variance explained by the regression model and can be converted (by multiplication by 100) to a percent of total variability explained by the specified linear combination of predictor variables. In this case, the value of R^2 is, by calculation as a ratio of sums of squared deviation scores, always non-negative. Its value always increases or remains stable with the addition of predictor variables into the model (the value of $R^2$is never reduced under these conditions) and can serve as a stand-alone measure of the predictive capability of a model.[1]

As multilevel modeling of data has become increasingly accessible to researchers, attention has been given to the need for a comparable measure of variance explained for these models. Because variance components can exist at each level of the multilevel model, the concept of "explained variance" becomes more complex. Several estimates may be needed for a single model, and R^2 may take on reduced,[2] or even negative, values with the addition of predictors. This portion of the chapter provides guidelines for calculating and interpreting these (sometimes anomalous) estimates and presents the two predominant methods.

The first method for computing and interpreting the multilevel model version of R^2, also sometimes referred to as *pseudo-R^2* (Singer & Willett, 2003), produces an R^2 statistic for each parameter estimate in the model (Raudenbush & Bryk, 2002). The statistic is interpreted as the *proportional reduction in variance* for that parameter estimate that results from the use of one model as compared to a base, or comparison, model. The statistic only can be computed and interpreted as the value of one model relative to another model and should not be interpreted as an explanation of the absolute amount of variance in the criterion variable.

The second method of deriving the multilevel R^2 statistic (Snijders & Bosker, 1994, 1999) results in separate measures of *proportional reduction in prediction error* for levels one (the prediction of Y_{ij}) and two (the prediction of $\overline{Y}_{.j}$) of the random intercepts only model. These estimates, too, represent changes in the amount of residual variance that result from the application of one model relative to a comparison model but make use of total estimated variance in their computation, as $\hat{\sigma}^2 + \hat{\tau}_{00}$ provides a reasonable estimate of the sample variance of Y (Snijders & Bosker, 1994).

In using R^2 with multilevel models, it is important to remain mindful of the unique interpretation of each estimate in drawing conclusions about model value.

R^2 as Proportional Reduction in Variance

At level one, the individual level, r_{ij} represents the random error associated with the measurement of individual i in group j (Singer & Willett, 2003) in relation to the estimated level-two group mean. Each level-one error is assumed to be normally distributed with a constant variance, represented as σ^2 (Raudenbush & Bryk, 2002). At the individual level, the proportional reduction in within-groups variance is calculated by first subtracting the level-one variance of the new model from that of the base model. The ratio of that difference to the base model level-one variance is interpreted "as a proportion reduction in that variance" (Kreft & deLeeuw 1998, p. 118). That statistic is computed

$$\frac{\hat{\sigma}_b^2 - \hat{\sigma}_f^2}{\hat{\sigma}_b^2} \qquad (7.6)$$

where $\hat{\sigma}_b^2$ = the estimated level-one variance for the base model and $\hat{\sigma}_f^2$ = the estimated level-one variance for the fitted model (Raudenbush & Bryk, 2002).

At level two, population variance components estimates are represented by $\hat{\tau}_{qq}$ and are given for the intercepts[3] (β_{0j}) and each slope estimate (β_{1j}, $\beta_{2j}, \ldots, \beta_{qj}$) that is not fixed to equal zero and, therefore, is permitted to be random. At level two, the proportional reduction in the variance of the intercepts, β_{0j}, is computed

$$\frac{\hat{\tau}_{00_b} - \hat{\tau}_{00_f}}{\hat{\tau}_{00_b}} \tag{7.7}$$

where $\hat{\tau}_{00_b}$ = the estimated variance of the intercepts in the base model and $\hat{\tau}_{00_f}$ = the estimated variance of the intercepts in the fitted model.

The result is a proportional reduction in variance that can be attributed to the predictor(s) unique to the fitted model (Raudenbush & Bryk, 2002).

Likewise, the proportional reduction in the variance of a given slope, β_{qj}, is calculated

$$\frac{\hat{\tau}_{qq_b} - \hat{\tau}_{qq_f}}{\hat{\tau}_{qq_b}} \tag{7.8}$$

where $\hat{\tau}_{qq_b}$ = the estimated variance of slope q in the base model and $\hat{\tau}_{qq_f}$ = the estimated variance of slope q in the fitted model.

Once a base model has been established, subsequent multilevel models can be compared to it to determine the resulting proportional reduction in variance. Note that level-two proportion reduction in variance statistics only can be compared for models with the same level-one model (Raudenbush & Bryk, 2002, p. 150). The estimated reduction is computed separately for each parameter estimate (Raudenbush & Bryk, 2002; Singer & Willett, 2003). For comparison of variance estimates at level one, the null model, which contains only a random intercept and no level-one predictor variables, typically is used as the base model, and other level-one models that include level-one predictors are compared to this null model. For comparison of variance estimates at level two, the level-one fitted model (containing level-one but no level-two predictor variables) is used as the base model for comparison of subsequent models containing level-two predictor variables.

Example

Table 7.2 contains variance components for several models fit to the ECLS-K data set, where the outcome is reading achievement of students in the fall of their kindergarten year. Model 1, which includes a level-one covariate, a measure of exposure to kindergarten (in months of instruction), serves as the base model. Model 2 includes an additional student-level predictor, (grand-mean centered) age at kindergarten entry. To determine the proportional reduction in level-one residual variance that resulted from the addition of the "age" predictor, the estimated student-level variances are compared according to Equation 7.6 above: (60.48 – 58.70)/60.48 = .029. This result indicates that only about 3% of the within-group variance, after accounting for months of kindergarten exposure, is attributable to student age at kindergarten entry. The bulk of the variance at this level is not explained by model 2.

The level-two variance estimates (τ_{00}) for the two models are compared as in Equation 7.7: (13.79 – 13.52)/13.79 = .019. As might be expected, the between-groups variance is reduced by a negligible amount (less than 2%) with the addition of the level-one predictor in model 2.

Model 5 (which includes *SES*, kindergarten exposure, and age of kindergarten entry as level-one predictors, and school sector and percent free lunch as level-two predictors) was compared to model 3 (a nested model that did not include the level-two predictors) to estimate proportional reduction in variance of the SES-achievement slopes (β_{1j}) using Equation 7.8: (2.19 – 2.17)/(2.19) = .01. We conclude that the level-two predictors, *school sector* and *percent free lunch*, in the fitted model do not explain any appreciable variance in the slope when compared to model 3.

The Confounding of Variance Estimates

As stated earlier, the concept of explained variance in multilevel models is different from that in OLS regression models as the former may involve multiple variance component estimates. An additional complexity of the concept in multilevel modeling relates to the confounding of variance component estimates across levels. We examine how the level-two variance component estimate in the random intercepts only model may be dependent on the specified level-one model.

When a significant level-one predictor is added to a random intercepts only model, the level-one variance component is reduced. This is to be expected. However, under these conditions, the level-two variance component, $\tau_{00,}$ also may be affected (Raudenbush & Bryk, 2002). Its value may increase or decrease with the addition of the level-one predictor. One explanation for the change is that, as level-one predictors are added to the random intercept models, the meaning of β_{0j} may change. Without any predictors in the model, this estimate represents the mean for group *j* on the outcome variable. As predictors are added, the interpretation of β_{0j} becomes the outcome for a person in group *j* whose value is zero for all level-one predictors (Raudenbush & Bryk, 2002). Unless all level-one variables are group mean centered, it is clear that this estimate, and its variance, $\tau_{00,}$ should change with the addition of variables in the level-one model. In the example above, the intercept for model 1 represents the average reading achievement for group *j* after adjusting for months of kindergarten exposure. In model 2, the intercept becomes the reading achievement for a student in school *j* whose entry age is equal to the grand-mean age of kindergarten entry because we have grand-mean centered the age variable.

In the random intercepts model, it is possible for level-two proportion reduction of variance statistics to be reduced or become negative with the addition of predictor variables at level one. This possibility conflicts with our interpretation of the traditional R^2 in OLS multiple regression as a sum of squares ratio and highlights another complexity in using an analog of R^2 in multilevel modeling. Essentially, because of the method by which variance components are estimated, while adding a level-two variable will decrease the estimate of τ_{00} and leave σ^2 relatively unchanged, adding a group-mean centered variable at level one will decrease the estimate of σ^2 *while increasing* the estimate of τ_{00} (Snijders & Bosker, 1994). When a fixed effect has been added to the model, Snijders and Bosker (1999) suggest that decreases in the proportion of variance explained of .05 or more in large sample studies also may be diagnostic of model misspecification.

To illustrate this phenomenon with the ECLS-K data set, we compared a model using group-mean centered SES at level one to the base model (model 1). The resulting estimate of σ^2 for the fitted model was 57.01 and for τ_{00} was 14.07. The proportion reduction in variance for level one was

$$\frac{\hat{\sigma}_b^2 - \hat{\sigma}_f^2}{\hat{\sigma}_b^2} = \frac{60.52 - 57.01}{60.52} = .06$$

where $\hat{\sigma}_b^2$ = the level-one variance for the base model and $\hat{\sigma}_f^2$ = the level-one variance for the fitted model. Group-mean centered SES explained approximately 6% of the within-groups variance.

At level two, the proportion of reduction in variance was

$$\frac{\hat{\tau}_{00_b} - \hat{\tau}_{00_f}}{\hat{\tau}_{00_b}} = \frac{13.79 - 14.07}{13.79} = -.02$$

where $\hat{\tau}_{00_b}$ = the level-two variance for the base model and $\hat{\tau}_{00_f}$ = the level-two variance for the fitted model. This illustrates that the inclusion of group-mean centered variables actually can increase variance at level two.

For comparison, we estimated the proportion of variance reduction for a model using *grand-mean centered* SES. For this model, $\hat{\sigma}^2$ was 57.28, and $\hat{\tau}_{00}$ was 5.58, resulting in a variance reduction estimate at level one of .05 and at level two of .60. Thus, where essentially no level-two variance was explained (in comparison to the base model) when the level-one predictor SES was group-mean centered, 60% of the variance was explained using grand-mean centering. A great deal of variability between schools on mean reading achievement can be accounted for by differences in student-level SES; however, within schools, this predictor explains a small proportion of the variance in achievement, indicating selection effects at the school level.

Generally speaking, group-mean centering will decrease σ^2 but increase τ_{00} (Snijders & Bosker, 1994). Therefore, the inclusion of a group-mean centered variable results in a reduction of unexplained variance at level one and an increase in unexplained variance at level two. For detailed discussion about the topic of centering variables and its effect on estimated variance reduction, we refer the reader to the existing literature (Enders & Tofighi, 2007; Hox, 2002; Kreft & deLeeuw, 1998; Raudenbush & Bryk, 2002; Snijders & Bosker, 1994, 1999).

R^2 as Proportional Reduction in Prediction Error

An alternative to estimating parameter-specific proportional reduction in variance that compensates for the confounding of variance estimates, R^2 can be computed as a *proportional reduction of prediction error.* This method of variance reduction estimation uses the total, rather than parameter-specific, variance estimates for each model in the comparison. R^2 is computed separately for levels one (the prediction of Y_{ij}) and two (the prediction of $\bar{Y}_{.j}$) (Snijders & Bosker, 1994, 1999). Given a random intercepts only model, the prediction error for individual outcomes (Y_{ij}) is equal to the sum of the level-one and level-two variance components,

$$\hat{\sigma}^2 + \hat{\tau}_{00}. \tag{7.9}$$

The proportional reduction of prediction error at level one for this model relative to the base or null model, R_1^2,

$$= 1 - \frac{\left(\hat{\sigma}^2 + \hat{\tau}_{00}\right)_f}{\left(\hat{\sigma}^2 + \hat{\tau}_{00}\right)_b} \tag{7.10}$$

where $(\sigma^2 + \tau_{00})_f$ = the prediction error for the fitted model and $(\sigma^2 + \tau_{00})_b$ = the prediction error for the base model.

Applying this formula to evaluate model 2 above relative to the base model (model 1), where $\hat{\sigma}^2 + \hat{\tau}_{00} = 74.27$ for model 1 and $\hat{\sigma}^2 + \hat{\tau}_{00} = 72.22$ for model 2, the *level-one* proportional reduction in prediction error, R_1^2, is 1 – (72.22/74.27) = .03

At level two, the prediction error for the group mean, $\bar{Y}_{.j}$,

$$= \frac{\hat{\sigma}^2}{n_j} + \hat{\tau}_{00}. \qquad (7.11)$$

where n_j represents the number of units in the level-2 cluster, j. When the numbers of units within a level-two cluster is unbalanced, there are a few options for the value of n_j (Hox, 2002; Snijders & Bosker, 1994). The researcher can determine a priori a value that is representative of all groups, use the average group size, or use the harmonic mean of the groups, calculated as $N/[\Sigma_j(1/n_j)]$ (Snijders & Bosker, 1999).

The level-two proportional reduction in the prediction error, R_2^2,

$$= 1 - \frac{\left(\dfrac{\hat{\sigma}^2}{n_j} + \tau_{00}\right)_f}{\left(\dfrac{\hat{\sigma}^2}{n_j} + \tau_{00}\right)_b} \qquad (7.12)$$

where

$$\left(\frac{\hat{\sigma}^2}{n_j} + \tau_{00}\right)_f$$

is the prediction error variance for the fitted model and

$$\left(\frac{\hat{\sigma}^2}{n_j} + \tau_{00}\right)_b$$

is the prediction error variance for the base model.

Applying Formula 12 to evaluate the change in variance at level two (comparing model 2 with model 1), given a representative value of $n = 12$,

$$R_2^2 = 1 - \frac{\dfrac{58.7}{12} + 13.52}{\dfrac{60.48}{12} + 13.79} = .02.$$

The relative size of values of R^2 at either level resulting from the two variance reduction estimation methods (whether they are similar or one estimate is larger than the other), depends on the effect of the predictor(s) in the model at that level. When a predictor significantly reduces the variance component at level one, and not at level two, the level-one proportion of reduction in variance estimate will be larger than the proportional reduction in prediction error variance at that level because the latter takes into account the small effect on the level-two variance. The opposite will be true for the level-two reduction in variance estimates, where the proportional reduction in prediction error variance will be the larger estimate at that level (because of the inclusion of the change to $\hat{\sigma}^2$ in that formula). The reverse relationships will occur when a predictor variable in a model results in a significant reduction in level-two variance and has little effect on the variance estimate for level one.

A benefit of the R^2-as-proportional reduction in prediction error is the (relatively) predictable "behavior" of R^2 under given conditions. When models are specified correctly and group size is constant at *n*, population values of R_1^2 and R_2^2 will be reduced when explanatory variables are removed, given the assumption that the variance components at levels one and two (r_{ij} and u_{0j}) are uncorrelated with all predictor variables, X_{ij} (Snijders & Bosker, 1994, 1999). However, sample estimates of R_1^2 and R_2^2 still may decrease when predictors are added or increase when predictors are deleted from the model (Snijders & Bosker, 1994, p. 355). Roberts and Monaco (2006) provide an example of the possibility of negative modeled variance using the proportional reduction in error variance model. A large reduction in value of R^2 with the addition of an explanatory variable into a model may be diagnostic of possible misspecification of the larger model (Snijders & Bosker, 1994). One "important type" of model misspecification is the restricting of predictor variables to have the same within- and between-group regression coefficients when they actually are different in the population (p. 356).

Estimating Variance Reduction in Three-Level Models

Snijders and Bosker (1999) provide a formula for estimating variance at level one of the three-level random intercept model. This simply involves adding the variance component from the third level to the formula for level-one variance in the two-level model (Formula 9), such that the total estimated variance equals the sum of the variance components at each level:

$$\hat{\sigma}^2 + \hat{\tau}_{00} + \hat{\varphi}_{00} \tag{7.13}$$

The level-one proportional reduction in residual variance then is calculated:

$$1 - \frac{(\hat{\sigma}^2 + \hat{\tau}_{00} + \hat{\varphi}_{00})_f}{(\hat{\sigma}^2 + \hat{\tau}_{00} + \hat{\varphi}_{00})_b} \tag{7.14}$$

where φ_{00} represents the variance component at level three (Snijders & Bosker, 1999).

Variance Reduction Estimates in Models with Random Slopes

In a model with random slopes, the relationship between the predictor variable and the dependent variable varies by cluster, and the variance component (τ_{qq}) estimates the variability in this relationship across clusters. According to Hox (2002), if the multilevel model contains random slopes, it is "inherently more complex, and the concept of explained variance has no unique definition" (p. 63). Estimating R_1^2 and R_2^2 for models with random slopes involves "tedious" calculation (see Snijders & Bosker, 1994). However, they can be estimated easily by omitting the random slopes and re-estimating "the models as random intercept models with the same fixed parts" (Snijders & Bosker, 1999, p. 105). The variance components from the fixed-slope model then should be used to calculate R_1^2 and R_2^2 as for random intercepts models, described earlier (Formulas 10 and 12). This typically results in estimates that are close approximations of those for the random slopes model (Snijders & Bosker, 1999).

Summary—Reduction in Variance Estimates

In conclusion, the various multilevel R^2-type statistics described above provide measures to compare one model to another in terms of its ability to account for the variability in a given data set. They are not, however, without their limitations, regardless of method of estimation. We briefly have presented two methods to estimate the proportional reduction in variance. Conclusions about the relative "value" of a model should be made carefully, with the unique definition of the variance reduction estimate in mind. When models have random slopes, R^2 does not have a unique definition (Hox, 1998; Kreft et al., 1995). The relationship between the level-one predictor and the dependent variable varies across level-two units, and the level-two variance estimate is not constant in these models (Snijders & Bosker, 1999). With these caveats in mind, multilevel R^2 measures can provide a useful tool to compare the predictive ability of various multilevel models.

CONCLUSION

Measures of model fit and model adequacy discussed in this chapter provide the researcher with fairly objective methods to compare multilevel models. However, researchers disagree about the appropriateness of these measures, and their use has been somewhat controversial. Therefore, it is important to use these measures thoughtfully and selectively, with attention to their limitations. Future methodological research should examine the properties and performance of these indices to define more clearly their utility to inform model selection in multilevel applications.

NOTES

1. Reporting adjusted R^2 in addition to R^2 is recommended when comparing regression equations with varying numbers of predictors. Adjusted R^2 is a corrected estimate of proportion of explained variance, accounting for sample size and number of predictors in the model (Green & Salkind, 2003).
2. Throughout this chapter, the symbol, R^2 is used to represent the proportional reduction in variance in the multilevel model. However, it should not be assumed that this statistic is the mathematical equivalent of, or analogous to, the squared multiple correlation R^2 used with OLS multiple regression.
3. The value and interpretation of the intercept, and variance of the intercept, at level two is influenced by the "location" of the Xvariable (i.e., the decision about whether and how X is centered) in level one (Raudenbush & Bryk, 2002).

REFERENCES

Box, G. E. (1979). Robustness in the strategy of scientific model building. In R. L. Lauer & G. N. Wilkinson (Eds.), *Robustness in Statistics* (pp. 201–236). New York: Academic Press.

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52,* 345–270.

Browne, M. W. (2000). Cross validation methods. *Journal of Mathematical Psychology, 44,* 108–132.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research, 33,* 261–304.

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the problem of sample size: A clarification. *Psychological Bulletin, 109,* 512–519.

de Leeuw, J. (2004). Multilevel analysis: Techniques and applications (Book review). *Journal of Educational Measurement, 41,* 73–77.

Enders, C. K. & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12,* 121–138.

Forster, M. R. (2000). Key concepts in model selection: Performance and generalizability. *Journal of Mathematical Psychology, 44,* 205–231.

Gelman, A., & Rubin, D. B. (1995). Avoiding model selection in Bayesian social research. *Sociological Methodology, 25,* 165–173.

Green, S. B., & Salkind, N. J. (2003). *Using SPSS for Windows and Macintosh: Analyzing and understanding data* (3rd ed.). Upper Saddle River, N.J.: Prentice Hall.

Gurka, M. J. (2006). Selecting the best linear mixed model under REML. *The American Statistician, 60, 1,* 19–26.

Hox, J. (2002). *Multilevel analysis techniques and applications.* Mahwah, N.J.: Lawrence Erlbaum Associates.

Kline, R. B. (1998). *Principles and practice of structural equation modeling.* New York: The Guilford Press.

Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research, 30,* 1–21.

Kreft, I. & de Leeuw, J. (1998). *Introducing multilevel modeling.* Thousand Oaks, CA: Sage Publications.

Kuha, J. (2004). AIC and BIC: Comparisons of assumptions and performance. *Sociological Methods and Research, 33,* 188–229.

O'Connell, A. A., Goldstein, J., Rogers, H. J., & Peng, C. Y. J. (2008). Logistic and ordinal multilevel models. In A. A. O'Connell and D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 199–242). Charlotte, NC: Information Age Publishing.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology, 25,* 111–163.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models.* Thousand Oaks, CA: Sage Publications.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. (2000). *HLM 5: Hierarchical linear and non-linear modeling.* Chicago: Scientific Software International.

Roberts, J. K., & Monaco, J. P. (2006, April). *Effect size measures for the two-level linear multilevel model.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461–464.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* Oxford: Oxford University Press.

Snijders, T., & Bosker, R. (1994). Modeled variance in two-level models. *Sociological Methods & Research, 22*(3), 342–363.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis.* Thousand Oaks, CA: Sage Publications.

Stapleton, L. M., & Thomas, S. L. (2008). The use of national datasets for teaching and research: Sources and issues. In A. A. O'Connell and D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 11–57). Charlotte, NC: Information Age Publishing.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data.* New York: Springer-Verlag.

Wagenmakers, E., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin and Review, 11,* 192–196.

Weaklim, D. L. (1999). A critique of the Bayesian Information Criterion for model selection. *Sociological Methods and Research, 27,* 359–397.

Weaklim, D. L. (2004). Introduction to the special issue on model selection. *Sociological Methods and Research, 33,* 167–187.

Whittaker, T. A., & Furlow, C. F. (2006, April). A comparison of model selection criteria for hierarchical linear modeling. Paper presented at the annual conference of the American Educational Research Association, San Francisco, CA.

Zucchini, W. (2000). An introduction to model selection. *Journal of Mathematical Psychology, 44*, 41–61.