A Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis

---

Leslie Rutwoski
Matthais von Davier
David Rutwoski

Chapter 1

Multilevel Modeling of Categorical Response Variables

Carolyn J. Anderson
University of Illinois, Urbana-Champaign

Jee-Seon Kim & Bryan S. B. Keller
University of Wisconsin, Madison

The most common type of item found on large-scale surveys has response options that are categorical. Models for binary, ordinal and nominal variables are relatively common and well developed (e.g., Agresti, 2002, 2007; McCullagh and Nelder, 1989; Powers and Xie, 1999; Fahrmeir and Tutz, 2001); however, three major complications arise when modeling responses from large-scale surveys that employ a complex sampling design. These challenges result from the fact that data are typically clustered, the probability of selection is not equal for all units, and missing data is the norm rather than the exception. To deal with the clustered or nested structure and to permit the investigation of context on individuals' responses, extensions of standard models for categorical data to clustered categorical data are presented. The particular random effects models presented in this chapter are logistic regression models for dichotomous responses, multinomial logistic regression models for nominal responses, and continuation ratios, adjacent categories, proportional odds and partially proportional odds models for ordinal data. The development of multilevel versions of these models and software to fit them has progressed to the point that the models are can be fit to data using a number of common software programs.

The second modification to the routine use of standard models for categorical data is dealing with unequal probability sampling of primary sampling units (e.g., schools) and secondary units (e.g., students). Ignoring the sample design can lead to biased results. Weighting of data may be necessary at each level of the model. The theory for adding weights to multilevel models is relatively straightforward (Pfeffermann et al., 1998; Asparouhov and Muthén, 2006; Grilli and Pratesi, 2004; Rabe-Hesketh and Skronkal, 2006); however, the availability of software for discrete data that implements the theory is much less common. Design weights are incorporated during estimation and the estimated parameters are based on maximizing the pseudolikelihood rather than maximizing the likelihood. The estimating equations for discrete models are presented in this chapter, because they open up alternative software options with minimal programming in statistical software programs.

The third issue when modeling data from large-scale surveys is the problem of missing data. Deleting cases that have missing data can lead to biased parameter estimates and is also very wasteful. Although numerous methods have been developed and are regularly used to deal with missing data, the options for missing data when data are hierarchically structured are limited. Multiple imputation for continuous (normal) data has been developed by Shin and Raudenbush (2007, 2010) and is described in Chapter **??**. Van Buuren (2011, 2012) presents a method based on fully conditionally specified models that can be used to impute missing values for both continuous and discrete variables. Unfortunately, this method has only been developed for imputing values at the lowest level

Table 1.1 *The distribution of response options to the item regarding use of the Internet to look up information for school.*

Response	Frequency	Percent	Cumulative percent
Every day or almost every day	746	14.94	14.94
Once or twice a week	1240	24.83	39.77
Once or twice a month	1377	27.57	67.34
Never or almost never	1631	32.66	100.00

of hierarchically structured data. At the possible cost of reduced efficiency, our approach to missing data imputes data at Level 2 and for each Level 1 unit thus preserving the multilevel structure.

In the first section of this chapter, the data used to illustrate the models and methodology is described along with how weights are computed and how missing data were imputed. In the following three sections, Sections 1.2, 1.3 and 1.4, models for dichotomous responses, nominal responses and ordinal responses, respectively, are presented along with example applications to the data described in Section 1.1. Current software options are described briefly in Section 1.5. The example input code for fitting models to data using Mplus and SAS and a short document explaining the input code can be downloaded from http://faculty.ed.uiuc.edu/cja/homepage. Lastly, we conclude in Section 1.6 with a summary of challenges in applications of models presented in this chapter.

1.1 The Data

The models presented in this chapter are illustrated using the United States data from the 2006 Progress in International Reading Literacy Study (PIRLS). The data are from a complex two-stage cluster sampling design with unequal probability weighting. Students are nested within classrooms (teachers) and classes are nested within schools. Although some schools have students from multiple classrooms, most schools have only one classroom in the study. Therefore, the data are treated as consisting of two levels: students within schools. There are $M = 182$ schools with a total of $N = 5,128$ students[1]. The number of students per school n_j ranges from 7 to 66 with an average of 27. Data from the fourth grade student questionnaire and the school questionnaire are used here.

1.1.1 Response and Explanatory variables

Given the increasing role of technology and the Internet in society, our goal is to study how various student and school factors are related to Internet usage, in particular the use of the Internet as a source of information for school related work. The response or "dependent" variable is an item asking students how often they use the Internet to "Look up information for school" (`Schnet`). The response options are "Every day or almost every day", "Once or twice a week", "Once or twice a month", and "Never or almost never". The distribution of the responses is given in Table 1.1.1. The response options are discrete with a natural ordering. All four categories are used to illustrate the nominal and ordinal models, and to illustrate models for binary logistic regression the responses are dichotomized as

$$Y_{ij} = \begin{cases} 1 & \text{If student } i \text{ in school } j \text{ uses the Internet at least once a week} \\ 0 & \text{If student } i \text{ in school } j \text{ uses the Internet at most twice a month} \end{cases}.$$

The within school or Level 1 explanatory variables include the students' gender (`Girl`), how

[1]This total reflects the fact that one school was excluded that had no responses on the school questionnaire, 9 students were deleted who has no responses on the student questionnaire, and 3 students were deleted who has missing data but the imputation method approach used failed due to too few students from their schools.

Table 1.2 *Basic summary statistics and information about explanatory variables used the examples. Values are based on non-missing data.*

Variable	Coding	Mean or percent	Std dev	Min	Max
Within or Level 1 Variables					
`Girl`	1 girl 0 boy	50.00% 50.00%			
`TimeRdg`	How much time per day spent reading for homework (ordinal but treated as quantitative)	2.34	.75	1	4
`ScreenTime`	Mean of 2 items regarding how much time per day spent watching TV, videos, electronic games (high score means more time)	2.34	1.13	0	4
Between or Level 2 Variables					
`NperComp`	The number of fourth grade students per computer (number)	3.05	3.30	.44	25.25
`Shortages`	Mean of 14 items dealing with various shortages (high score means more shortages)	0.59	.54	0	3.0
`Urban`	1 Urban school 0 Suburban or rural	29.44% 70.65%			
`Suburban`	1 Suburban school 0 Urban or rural	42.78% 57.22%			
`Rural`	1 Rural school 0 Urban or suburan	27.78% 72.22%			
`AllFree`	1 All students free/reduced price lunch 0 Some or none	11.67% 83.33%			
`SomeFree`	1 Some students free/reduced price lunch 0 All or none	79.44% 20.56%			
`NoneFree`	1 Few students free/reduced price lunch 0 All or some	8.89% 91.11%			

much time a student spends reading for homework (`TimeRdg`), and the amount of time per day a student spends watching TV, videos, playing computer games, etc. (`ScreenTime`). The between school or Level 2 explanatory variables include the number of fourth grade students per computer designated for fourth grade student use (`NperComp`), shortages of materials and staff at the school (`Shortages`), the location of the school (`Urban`, `Suburban`, `Rural`), and whether some, all or no students at a school receive free or reduced price lunches (`AllFree`, `SomeFree`, `None`). To ensure the proper interpretation of model results, more information about these variables and basic summary statistics are given in Table 1.1.1.

1.1.2 Weights

The recommendations for weights given by Rutkowski et al. (2010) are used here. The weights for students (Level 1 or secondary units) will be computed as the product of student and class weights:

$$w_{1|ij\ell} = \underbrace{(WF_{ij\ell} \times WA_{ij\ell})}_{student\ i}\ \underbrace{(WF_{j\ell} \times WA_j)}_{class\ell}, \tag{1.1}$$

where $WF_{ij\ell}$ is the inverse of the probability of selection of student i and class ℓ from school j, $WF_{j\ell}$ is the inverse of the probability of selection of class ℓ from school j, $WA_{ij\ell}$ is the weight adjustment for non-response for student i in class ℓ from school j, and $WA_{j\ell}$ is the weight adjustment for non-response for class ℓ from school j. The weight adjustments are for those students and classes that were selected but chose not to participate (Heeringa et al., 2010; Rutkowski et al., 2010). The school weights used are

$$w_{2|j} = WF_j \times WA_j, \tag{1.2}$$

where WF_j is the inverse of the probability of selecting school j and WA_j is the weight adjustment for school j for non-response.

There are two opinions on the use of sampling or design weights in an analysis. One approach is design-based and advocates using sampling weights in any analysis of the data. The other approach is model-based and advocates not using the sampling weights. Snijders and Bosker (2012) discusses this issue in the context of multilevel models. The main argument in favor of a model-based approach is that the sample design is irrelevant when the model is the "true" one and the sampling procedure is independent of the probability model. If this is the case, taking into account sampling weights results in a loss of efficiency and less precise estimates (Heeringa et al., 2010). The main argument for a design-based approach is that parameter estimates could be seriously biased, especially if the sampling weights vary widely over units in the population (i.e., the weights are informative). Since there is a trade-off between efficiency (model-based has better efficiency) and bias (design-based is unbiased), both model and design based results are reported.

To determine whether the design may influence the results, one can examine the variability of the weights, the effective sample size, and the design effect. For example, if the schools had the same probability of being selected and every school selected responds, then the weights for the schools would all be equal. When weights are all the same, their variance equals 0 and they could be simply set to 1. With equal weights, the design would be simple random sampling. The same is true for students. If each student has the same probably of selection and each student responds, then the variance of the weights would be 0. For the PIRLS data, the mean and standard deviation of the weights for schools equal 305.42 and 218.64, respectively, which suggest the school weights are informative. For the students, the means of the Level 1 weights of students within schools mostly equal 1, and for 74% of the schools, the standard deviations equal 0. The standard deviations for the other 26% of the schools are less than 0.13 and most (i.e., 21%) are less than 0.05. The relatively large value of the standard deviation of the school weights suggests that the school weights will have an impact on the results; however, the small standard deviations for the student weights suggests these will have a negligible impact on the results.

Another way to assess the potential impact of weights is to examine the effective sample sizes. The effective sample sizes for the Level 2 (primary) and Level 1 (secondary) units are defined as

$$N_{\text{effective}} = \frac{(\sum_j w_{2|j})^2}{\sum_j (w_{2|j}^2)} \quad \text{and} \quad n_{\text{effective},j} = \frac{(\sum_i \sum_\ell w_{1|ij\ell})^2}{\sum_i \sum_\ell (w_{1|ij\ell}^2)}, \tag{1.3}$$

respectively (Heeringa et al., 2010; Snijders and Bosker, 2012). These formulas define an effective sample size such that the information in a weighted sample equals the effective sample size from simple random sampling Snijders and Bosker (2012). If the weights are all equal, then the effective sample sizes would equal the actual sample size. If the weights are informative, then the effective

sample size will be less than the actual sample size. For the PIRLS data, the effective sample size for the schools is only 120, which is considerably less than the number of schools (i.e., $M = 182$). The effective sample sizes for students within schools mostly equal n_j, the number of students from school j.

A third measure of impact of the sampling design is based on the effective sample size. The design effects for Level 2 and Level 1 are defined as the ratios of the effective sample size over the actual sample size,

$$\text{Design} = N_{\text{effective}}/M \quad \text{and} \quad \text{Design}_j = n_{\text{effective},j}/n_j, \tag{1.4}$$

where M equals the number of clusters (Snijders and Bosker, 2012). If weights are non-informative, then effective sample size equals the actual sample size and the ratio of the effective over the actual sample size equals 1. For the PIRLS data, the Level 2 design effect only equals 0.66; however, the Level 1 design effects are all mostly equal to one (i.e., $\text{Design}_j = 1$). In the analysis of our data, the Level 1 weights will have a negligible effect on the results, but the Level 2 weights will likely have a noticeable impact.

The last issue related to weights is how to scale them. In the context of multilevel models, two major suggestions are given for scaling of the weights (Pfeffermann et al., 1998): sums of weights equals effective sample size or sum of weights equals sample size. In the example analyses using the PIRLS data, we use the more common approach and scale weights such that their sums equal the sample sizes (i.e., $\sum_j w_{2|j} = 182$ and $\sum_i \sum_\ell w_{1|ij\ell} = n_j$).

The weights are incorporated when estimating a model's parameters. How weights are incorporated is described in detail in Section 1.2.3 for binary logistic regression model and is subsequently modified for the nominal and ordinal models in Sections 1.3 and 1.4. The capability to include weights for the models discussed in this chapter has not been implemented in many of the common programs. For the examples presented in this chapter, SAS (version 9.3) was used to fit all models to the PIRLS data and Mplus (Muthén and Muthén, 1998-2010) was also used for the models that Mplus is capable of fitting to data (i.e., random effects binary logistic regression and proportional odds models). Software options are discussed in more detail in Section 1.5.

1.1.3 Missing Data

A vexing problem for large, complex surveys (and longitudinal data) is missing data. Dealing with missing data is particularly difficult when data are clustered. If students or schools with missing data on the variables listed in Table 1.1.1 were excluded, there would only be $M = 158$ schools and a total of $3,994$ students. Simply removing students or whole schools due to missing data is not only a waste of data, but also can result in biased parameter estimates (Allison, 2002; Schafer, 1997; Van Buuren, 2011, 2012; Enders, 2010). The program Mplus can fit models using maximum likelihood estimation with missing values on the response variable, but cannot handle missing predictors. Multiple imputation is an alternative approach to impute missing response and predictor variables.

A few procedures have been developed for multiply imputing missing clustered data. Chapter **??** of this book deals with missing data in multilevel models; however, this only pertains to normally distributed variables (Shin and Raudenbush, 2007, 2010). Another proposal for multilevel data that is described by Van Buuren (2011) (see also Van Buuren, 2012) only deals with simple cases without missing Level 2 variables. Neither of these two solutions work for our data. Yet another proposal have been put forth that includes dummy variables for each cluster (Reiter et al., 2006); however, this presupposes that clusters only differ in terms of their intercepts. In our example, we do not want to make this assumption and using dummy variables is impractical due to the large number of clusters; therefore, this method is not pursued here. Nearly non-existent are proposals for incorporating sampling weights into the imputation model for missing data. The one exception is Amer (2009), but this only deals with two clusters.

Ignoring the nested structure of the data when multiply imputing missing values can lead to

Table 1.3: *Missing data on student questionnaire.*

Pattern	Girl	Schnet	TimeRdg	ScreenTime	Frequency	Percent
1	X	X	X	X	4715	91.95
2	X	X	X	.	219	4.27
3	X	.	X	X	69	1.35
4	X	.	X	.	51	0.99
5	X	X	.	X	48	0.94
6	X	X	.	.	7	0.14
7	X	.	.	X	12	0.23
8	X	.	.	.	4	0.08
9	.	X	X	X	3	0.06
Frequency	3	136	71	281	5128	
Percent	0.06	2.6548	1.38	5.48		

severely biased results (Reiter et al., 2006; Van Buuren, 2011, 2012); therefore, we took a practical approach. Since the PIRLS data are clustered, the imputations were carried out separately by imputing school level missing values (i.e., one analysis on the 182 schools) and imputing missing values for student level. The latter was done by carrying out 182 analyses, one for each school. Although this method is not optimally efficient (leading to perhaps overly conservative inference), this approach retains associations among school (Level 2) variables, and preserves the heterogeneity between schools (i.e., random differences between schools in terms of intercept and the effects of variables). Furthermore, since the Level 1 weights are non-informative, these have minimal impact on the quality of the Level 1 imputed data.

For the PIRLS data, a reasonable assumption is that the data are missing at random, and we included additional variables in the imputation model to increase the likelihood of meeting this assumption. The patterns of missing data in the current study for the student variables are given in Table 1.1.3 and those for the school variables are given in Table 1.1.3. Each row represents a pattern of missing data where an "X" indicates that the column variable is not missing and a "." indicates that the variable is missing. The last two columns give the frequency and percent that the row patterns occur, and the last two rows give the frequency and percent of missing values for each variable. The percent of students that have no missing data is fairly high (i.e., 92%), and the percentages of missing values for each variable are relatively small (i.e., less than 5.5%). In Table 1.1.3, two school level variables that comprise `StdComp` are given (i.e., `Num4th`, the number of fourth grade students, and `SchComp`, the number of computers available to fourth grade students). From Table 1.1.3, we find that 84% of the schools' data are completely observed and most school level variables have less than 5% missing values. The one exception is `SchComp` that has 7.7% of the values missing.

Fully conditionally specified models implemented in SAS (versions 9.3) were used to impute missing school and student level variables (Van Buuren, 2011, 2012). Fifteen imputed data sets were created. The relative efficiency was examined for each imputed variable to determine whether fifteen data sets were sufficient. The relative efficiency measures the multiple imputed standard error relative to it's theoretical minimum value, more specifically it equals

$$\left(1 - \frac{\text{Fraction of missing information}}{\text{Number of imputered data sets}}\right)^{-1}$$

(Enders, 2010). The imputation models included additional auxiliary variables from the student and school questionnaires and imputations were done at the scale level (rather than at the item level). For the school questionnaire data, 9 variables were used in the imputation model, 4 of which were auxiliary variables. The relative efficiencies were all larger than .99. For the student level data, the imputation model was kept relatively simple (i.e., only 4 variables, one of which was an auxiliary

Table 1.4: *Missing data on school questionnaire.*

Group	Lunch	Location	Num4th	Short	Schcomp	Frequency	Percent
1	X	X	X	X	X	152	83.52
2	X	X	X	X	.	11	6.04
3	X	X	X	.	X	6	3.30
4	X	X	X	.	.	2	1.10
5	X	X	.	X	X	6	3.30
6	X	X	.	.	.	1	0.55
7	X	.	X	X	X	2	1.10
8	.	X	X	X	X	2	1.10
Frequency	2	2	7	9	14	182	
Percent	1.10	1.10	3.85	4.95	7.69		

variable) due to small sample sizes in some schools. When imputing the student level data, only 2 of the 182 schools were problematic and these had small sample numbers of students. The three students from these two schools (1 from one school and 2 from the other) were dropped. The mean relative efficiencies for the within school imputations were all greater than .99. The final number of students in the data set equaled 5,128.

After the data were imputed, the school and student data sets were combined, composite and dummy variables that are used in the analyses were created, and the design weights were computed and scaled.

1.2 Dichotomous Response Variables

When a response variable is dichotomous, natural and common choices to model the responses are logistic and probit regression models, and when respondents are nested or clustered within larger units, multilevel random effects versions of these models. In this chapter we focus on the multilevel logistic regression model.

In Section 1.2.1, the multilevel random effects logistic regression model is presented as a statistical model for clustered data, and in Section 1.2.2, the model is presented in terms of a latent variable, including random utility formulation. In Section 1.2.3, estimation that incorporates design weights is discussed. In Section 1.2.4, an analysis of the PIRLS data is presented to illustrate the model.

1.2.1 Multilevel Binary Logistic Regression

A single level logistic regression model can be generalized to a multilevel model by allowing the regression coefficients to differ randomly over clusters. Let the response variable for individual i in group j be coded as $Y_{ij} = 1$ for responses in a target category and $Y_{ij} = 0$ for responses in the other category. The distribution of Y_{ij} is assumed to be Bernoulli (or Binomial if data are tabulated). A cluster-specific or Level 1 logistic regression model for the probability that individual i in cluster j has a response in the target category (i.e., $P(Y_{ij} = 1)$) is

$$P(Y_{ij} = 1) = \frac{\exp(\beta_{0j} + \beta_{1j}x_{1ij} + \ldots + \beta_{Pj}x_{Pij})}{1 + \exp(\beta_{0j} + \beta_{1j}x_{1ij} + \ldots + \beta_{Pj}x_{Pij})}, \tag{1.5}$$

where x_{pij} is the value of the p^{th} Level 1 predictor variable for individual i in cluster j, β_{0j} is the intercept for cluster j, and β_{pj} is the regression coefficient for x_{pij} in cluster j. Note that for $P(Y_{ij} = 0)$, the β_{pj}'s in the numerator all equal zero. The term in the denominator ensures that $P(Y_{ij} = 1) + P(Y_{ij} = 0) = 1$.

The between cluster or Level 2 models are the same as those in the multilevel models for normal response variables discussed in Chapter **??** (i.e., hierarchical linear models or HLMs). Suppose that there are Q possible between cluster predictors, $z_{1j}, \ldots, z_{Qj}$. The Level 2 models are linear models for each Level 1 regression coefficient:

$$\begin{aligned}
\beta_{0j} &= \sum_{q=0}^{Q} \gamma_{0q} z_{qj} + U_{0j} \\
\beta_{1j} &= \sum_{q=0}^{Q} \gamma_{1q} z_{qj} + U_{1j} \\
&\vdots \\
\beta_{Pj} &= \sum_{q=0}^{Q} \gamma_{Pq} z_{qj} + U_{Pj},
\end{aligned} \tag{1.6}$$

where $z_{0j} = 1$ for an intercept. The z_{qj}s are predictors or explanatory variables that model systematic differences between clusters, the γ_{pq}s are fixed regression coefficients for the z_{qj}s, and the U_{pj}s are the cluster-specific random effects. The first subscript on γ_{pq} corresponds to the effect in the Level 1 model and the second subscript corresponds to the predictor in the Level 2 model. For example, γ_{pq} is the regression coefficient for z_{qj} in the model for β_{pj}.

The predictors in the models for the β_{pj}s (i.e., the z_{qj}s) need not be the same over the models in (1.6). For example, in the PIRLS data, shortages of supplies may predict differences between schools' intercepts, but not predict the differences between schools in terms of the effect of the amount of time that a student spends using electronics (i.e., `ScreenTime`).

The unexplained, random or stochastic differences between clusters are modeled by the random effects. The distributional assumption for the U_{pj}s is

$$\boldsymbol{U}_j = \begin{pmatrix} U_{0j} \\ U_{1j} \\ \vdots \\ U_{Pj} \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{10} & \ldots & \tau_{P0} \\ \tau_{10} & \tau_{11} & \ldots & \tau_{P1} \\ \vdots & \vdots & \ddots & \vdots \\ \tau_{P0} & \tau_{P1} & \ldots & \tau_{PP} \end{pmatrix} \right) \; i.i.d., \tag{1.7}$$

where MVN stands for multivariate normal and $i.i.d$ stands for independent and identically distributed. For short, assumption 1.7 can be written as $\boldsymbol{U}_j \sim MVN(\boldsymbol{0}, \boldsymbol{T})$ $i.i.d.$.

Substituting the Level 2 models into the regression coefficients of the Level 1 model yields our combined or cluster-specific model for the probabilities,

$$P(Y_{ij} = 1 | \boldsymbol{x}_{ij}, \boldsymbol{z}_j, \boldsymbol{U}_j) = \frac{\exp(\sum_{p=0}^{P}(\sum_{q=0}^{Q} \gamma_{pq} z_{qj} + U_{pj}) x_{pij})}{1 + \exp(\sum_{p=0}^{P}(\sum_{q=0}^{Q} \gamma_{pq} z_{qj} + U_{pj}) x_{pij})}, \tag{1.8}$$

where $x_{0ij} = z_{0j} = 1$ for the intercept. To emphasize that these are conditional models for probabilities that depend on the observed and unobserved variables, the conditioning is explicitly indicated here where $\boldsymbol{x}_{ij}$ consists of within cluster predictors, $\boldsymbol{z}_j$ the observed between cluster predictors, and $\boldsymbol{U}_j$ the unobserved between cluster random effects.

Similar to single level logistic regression models, $\exp(\gamma_{pq})$ equals an odds ratio. More specifically, based on the model for probabilities in (1.8), the odds that $Y_{ij} = 1$ versus $Y_{ij} = 0$ equals

$$\frac{P(Y_{ij} = 1 | \boldsymbol{x}_{ij}, \boldsymbol{z}_j, \boldsymbol{U}_j)}{P(Y_{ij} = 0 | \boldsymbol{x}_{ij}, \boldsymbol{z}_j, \boldsymbol{U}_j)} = \exp\left[\sum_{p=0}^{P} (\sum_{q=0}^{Q} \gamma_{pq} z_{qj} + U_{pj}) x_{pij} \right]. \tag{1.9}$$

To illustrate the interpretation of the γ_{pq}s, consider a model with two Level 1 (within cluster) predictors, x_{1ij} and x_{2ij}, one Level 2 (between group) predictor z_{1j} of the intercept and the slope of x_{2ij},

and a random intercept U_{0j}. The odds for this model are

$$\frac{P(Y_{ij}=1|x_{1ij},x_{2ij},z_{1j},U_{0j})}{P(Y_{ij}=0|x_{1ij},x_{2ij},z_{1j},U_{0j})}=\exp(\gamma_{00}+\gamma_{01}z_{1ij}+\gamma_{10}x_{1ij}+\gamma_{20}x_{2ij}++\gamma_{21}z_{1j}x_{2ij}+U_{0j}). \tag{1.10}$$

To interpret the parameter of a "main effect", for example the effect of x_{1ij}, holding *all other variables constant*, the odds that $Y_{ij}=1$ for a one unit increase in x_{1ij} is $\exp(\gamma_{10})$ times the odds for x_{1ij}; that is, the odds ratio equals $\exp(\gamma_{10})$. If $\gamma_{10}>0$ then the odds increase, if $\gamma_{10}<0$ then the odds decrease, and if $\gamma_{10}=0$ then the odds do not change (i.e., the odds are equal). This interpretation of the effect of x_{1ij} requires that the values of all other variables be constant; however, it does not depend on the particular values of the other predictor variables or the particular value of x_{1ij}. To interpret a (cross-level) interaction when the focus is on the effect of x_{2ij} on Y_{ij}, we would consider the odds for a one unit increase in x_{2ij}. This odds ratio is $\exp(\gamma_{20}+\gamma_{21}z_{1j})$ and it depends on the value of z_{1j}. In reporting and explaining the effect of an interaction, representative values of z_{1j} could be used (e.g., 25^{th}, 50^{th} and 75^{th} percentiles). Alternatively, when the focus is on the effect of z_{1j} on Y_{ij}, we could report and explain the odds ratio for a one unit increase in z_{ij}, which is $\exp(\gamma_{01}+\gamma_{21}x_{2ij})$.

The interpretation of γ_{pq} is always qualified by "all other variables constant", including the random effects U_{0j} and other observed predictors. As a result the interpretation is cluster-specific or within clusters because the random school effect U_{0j} has to be held constant. This is different from the interpretation of γ_{pq}s in HLM. The γ_{pq}s in HLM are interpretable as cluster-specific and as marginal or population average effects. This difference stems from the fact that marginal distributions of multivariate normal random variables (as we have in HLM) are normally distributed. This is not true for a multilevel random effects logistic regression model or any of the models covered in this chapter. When one collapses over the unobserved random effects of the cluster-specific logistic regression model to get the marginal distribution, the result is not a logistic regression model (Demidenko, 2004; Raudenbush and Bryk, 2002; Snijders and Bosker, 2012). The marginal effects in multilevel random effects logistic regression are smaller than the cluster specific ones.

1.2.2 A Latent Variable Approach

An alternative approach to random effects models for dichotomous responses (and later multicategory responses) is to propose a latent continuous variable that underlies the observed data (Muthén, 1998–2004; Skrondal and Rabe-Hesketh, 2000). This approach can also be framed as a random utility model or a discrete choice model (McFadden, 1974). The former hypothesizes that observed data are related to the latent variable Y^*_{ij} as follows:

$$Y_{ij}=\begin{cases}1 & \text{if } Y^*_{ij}>0\\ 0 & \text{if } Y^*_{ij}\leq 0\end{cases}, \tag{1.11}$$

where 0 is the threshold. A linear random effects model is then proposed for the latent variable Y^*_{ij},

$$Y^*_{ij}=\sum_{p=0}^{P}\left(\sum_{q=0}^{Q}\gamma_{pq}z_{pqk}+U_{pj}\right)x_{pij}+\varepsilon_{ij}, \tag{1.12}$$

where $z_{00k}=x_{0ij}=1$ (for intercepts), γ_{pq} are the fixed effects parameters, U_{pj} are the random effects for clusters, and ε_{ij} is a random residual for individual i within cluster j. This linear model can also be arrived at using a multilevel perspective as was done in the previous section.

The distribution assumed for ε_{ij} determines the model for data. If $\varepsilon_{ij}\sim N(0,\sigma^2)$, the model for Y_{ij} is a probit model, and if ε_{ij} follows a logistic distribution, the model for Y_{ij} is a logistic regression[2]

[2]In the random utility or choice model formulation, a linear model is given for the latent variable (utility) for each

Instead of 0 as the cut-off or threshold relating the observed and latent variables in (1.11), some authors and programs (e.g., Muthén & Muthén in Mplus) estimate a non-zero threshold, say ξ. This is equivalent to using a 0 cutoff by noting that for $Y_{ij} = 1$,

$$Y^*_{ij} = \gamma_{00} + \sum_{q=1}^{Q} \gamma_{0q} z_{0qk} + U_{0j} + \sum_{p=1}^{P} (\sum_{q=0}^{Q} \gamma_{pq} z_{pqk} + U_{pj}) x_{pij} + \varepsilon_{ij} > 0,$$

which is equivalent to

$$\sum_{q=1}^{Q} \gamma_{0q} z_{0qk} + U_{0j} + \sum_{p=1}^{P} (\sum_{q=0}^{Q} \gamma_{pq} z_{pqk} + U_{pj}) x_{pij} + \varepsilon_{ij} > -\gamma_{00}.$$

The cut-off or threshold is $\xi = -\gamma_{00}$.

The latent variable approach facilitates the extension of the model to multi-category data. In the case of ordered response variables, assuming the existence of a latent variable is especially appealing because there may be some underlying continuum that gives rise to ordered observed responses. Many psychometric models are based on just such an assumption (e.g., random utility models, discrete choice models, Guttman scale, item response theory models, and others).

Similar to HLM for normal response variables, an intra-class correlation (ICC) can be computed for random intercept models. In the case of a random intercept logistic regression model with no predictors (i.e., the only random effect in the model is U_{0j}), the ICC equals

$$ICC = \frac{\tau_{00}}{\tau_{00} + \pi^2/3}.$$

The ICC is a measure of within cluster homogeneity and equals the proportion of variance due to between cluster differences. A *residual* ICC, computed when there are fixed effect predictors and a random intercept, measures within cluster homogeneity and variance due to between cluster differences given the predictors.

1.2.3 Estimation

Typically, maximum likelihood estimation (MLE) is used to estimate single level and multilevel random effects logistic regression model parameters; however, with weights, the method used is a version of pseudolikelihood estimation. Following the general approach of Pfeffermann et al. (1998) (see also, Grilli and Pratesi, 2004; Asparouhov and Muthén, 2006; Rabe-Hesketh and Skronkal, 2006), the weights are incorporated into the likelihood equations and the parameters are found by maximizing the modified equations or pseudolikelihood equations. The basic procedure described below is also used for other models covered in this chapter with slight variations (i.e., the distribution of cluster-specific response and the model for probabilities). Given the multilevel structure, first we show how the Level 1 weights are included to yield cluster-specific pseudolikelihood equations and then show how the Level 2 weights are incorporated.

Let $\mathcal{L}(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{z}_j, \boldsymbol{U}_j)$ equal the logarithm of the likelihood conditional on the observed predictor variables and random effects $\boldsymbol{U}_j$ for individual i in cluster j. To incorporate the Level 1 weights $w_{1|ij}$, each of the likelihoods for individuals within a cluster are weighted by their values of $w_{1|ij}$ and then the random effects are collapsed over to yield the cluster-specific log-pseudolikelihood $\mathcal{L}(\boldsymbol{y}_j)$ as follows:

$$\mathcal{L}(\boldsymbol{y}_j) = \log \left[\int_{\boldsymbol{U}} \exp \left\{ \sum_{i=1}^{n_j} w_{1|ij} \mathcal{L}(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{z}_j, \boldsymbol{U}_j) \right\} f(\boldsymbol{U}_j) d\boldsymbol{U}_j \right], \tag{1.13}$$

category (i.e., Y^*_{ijk} for category k where $k = 0, \ldots, K$), and the category with the largest value of the latent variable/utililty is selected. For identification, the random and fixed effects for the utility associated with the reference category (i.e., $Y^*_{ij0} = 0$) equal zero; therefore, Y^*_{ij} in (1.11) represents the difference between utilities. The distribution of the difference between ε_{ij}s determines the choice option selected. When the residuals in the model for Y^*_{ijk} are assumed to follow a Gumbel (extreme value) distribution, the difference between the ε_{ij}s follows a logistic distribution leading to the logistic regression model.

where integration is over all random effects in $\boldsymbol{U}_j$, and $f(\boldsymbol{U}_j)$ is the multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{T}$ (Asparouhov and Muthén, 2006; Rabe-Hesketh and Skronkal, 2006; Grilli and Pratesi, 2004). In (1.13), the unobserved predictors are integrated out to yield cluster-specific log-pseudolikelihoods. For logistic regression, the individual likelihood equals the Bernouli distribution function and $\mathcal{L}(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{z}_j, \boldsymbol{U}_j)$ is

$$\mathcal{L}(y_{ij}|\boldsymbol{x}_{ij}, \boldsymbol{z}_j, \boldsymbol{U}_j) = \log\left\{P(Y_{ij} = 1|\boldsymbol{x}_{ij}, \boldsymbol{z}_j, \boldsymbol{U}_j)^{y_{ij}}(1 - P(Y_{ij} = 1|\boldsymbol{x}_{ij}, \boldsymbol{z}_j, \boldsymbol{U}_j))^{(1-y_{ij})}\right\},$$

where $P(Y_{ij} = 1|\boldsymbol{x}_{ij}, \boldsymbol{z}_j, \boldsymbol{U}_j)$ is given by (1.8), and y_{ij} is an indicator variable that equals 1 if the response by individual i in cluster j is the target category and 0 otherwise.

The group specific log-pseudolikelihoods in (1.13) must be combined to yield the log-pseudolikelihood using all the data. The Level 2 weights enter at this point. Assuming independence between clusters, the log-pseudolikelihood for all the responses equals

$$\mathcal{L}(\boldsymbol{y}) = \sum_{j=1}^{M} w_{2|j}\mathcal{L}(\boldsymbol{y}_j). \tag{1.14}$$

(Grilli and Pratesi, 2004; Asparouhov and Muthén, 2006; Rabe-Hesketh and Skronkal, 2006). The parameters (i.e., γ_{pq} and $\tau_{pp'}$ for all p, p' and q) that maximize (1.14) are the maximum pseudo-likelihood estimates. Although, the log-pseudolikelihood equations have only been given here for a two level models, Rabe-Hesketh and Skronkal (2006) give a general set of formulas for higher-level models and illustrate their use for a 3–level model fit to PISA data.

A number of estimation algorithms exists that attempt to find the parameters that maximize either the log-likelihood (i.e., $w_{2|j} = w_{1|ij} = 1$) or the log-pseudolikelihood. Two algorithms, marginal quasi-likelihood and penalized quasi-likelihood, attempt to approximate the model by linearizing the model using a Taylor series expansion and then using an algorithm designed for a linear mixed model (SAS Institue Inc, 2011b). Unfortunately, these two strategies yield parameters estimates that are severely biased. A better approach is to find the parameters that maximize the function in (1.14). The "gold standard" is adaptive Gaussian quadrature (i.e., numerical integration); however, this becomes computationally very difficult and time consuming for multiple (correlated) random effects. A third alternative is Bayesian methods. For the examples presented in this chapter, adaptive quadrature is used to estimate model parameters.

When MLE is used to obtain parameter estimates, the standard errors of parameters can be estimated based on the model and are valid provided that the correct model is specified. Under pseudo-likelihood estimation, the model based standard errors will be biased; therefore, robust (sandwich or empirical) estimators of the standard errors are recommended. The sandwich estimates are based on the data[3]. The robust standard errors can be used to compute test statistics for the fixed effects, in particular, the test statistic equals $\hat{\gamma}_{pq}/\widehat{se}$ where $\hat{\gamma}_{pq}$ is the pseudolikelihood estimate of γ_{pq} and $\widehat{se}$ is the sandwich estimate.

1.2.4 Example for Binary Response Variable

In this example, we model how much time a student spends looking up information for school on the Internet where the response variable was coded $Y_{ij} = 1$ for at least once a week and $Y_{ij} = 0$ for at most twice a month. The predictor variables that were considered are given in Table 1.1.1. Each model was fit to each of the 15 imputed data sets and the results combined using Rubin's method (Rubin, 1987), which is given in Chapter **??** of this book. As a measure of the impact of missing data on the analysis, the missing fraction of information for estimating a parameter was computed. This measures how much information about a parameter is lost due to missingness (Snijders and

[3]The sandwich estimators are also used in MLE with a misspecified model.

Bosker, 2012). This fraction equals

$$\text{Missing Fraction} = \frac{(1+1/15)\text{var}(\hat{\gamma}_{pq})}{\widehat{\text{var}}(\hat{\gamma}_{pq})+(1+1/15)\text{var}(\hat{\gamma}_{pq})}, \tag{1.15}$$

where $\widehat{\text{var}}(\hat{\gamma}_{pq})$ is the estimated sandwich variance from combining over imputations and $\text{var}(\hat{\gamma}_{pq})$ is the variance of the parameters over the 15 imputed data sets. Small values indicate little loss of information.

The models were fit using both Mplus (version 6) and SAS PROC NLMIXED (verison 9.3) and empirical (sandwich) standard errors were computed by each program. We started with a relatively simple random intercept model with few fixed effects and increased the complexity by adding more fixed effects at Level 1 and Level 2, including cross-level interactions. Some of the Level 1 predictor variables showed a fair amount of variability between schools; therefore, Level 1 variables were centered around their school means and the means used as predictors of the intercept. The first such model had school-mean centered `ScreenTime` (i.e., $\texttt{ScreenTime}_{ij} - \overline{\texttt{ScreenTime}}_j$) and school-mean centered `TimeRdg` (i.e., $\texttt{TimeRdg}_{ij} - \overline{\texttt{TimeRdg}}_j$) as within or Level 1 predictors, and the means $\overline{\texttt{ScreenTime}}_j$ and $\overline{\texttt{TimeRdg}}_j$ were entered into the model as predictors of school intercepts. Once a relatively complex fixed effects model was successfully fit to the data, fixed effects that were non-significant were successively dropped; however, at some steps some school effects were put back in the model and re-tested. After arriving at a fixed effects structure that seemed to be best, we fit three more models each of which had a random slope for each of the Level 1 variables. The final model chosen was

$$\text{Level 1:} \quad \log\left(\frac{P(Y_{ij}=1)}{P(Y_{ij}=0)}\right) = \beta_{0j} + \beta_{1j}\texttt{Girl}_{ij} + \beta_{2j}\texttt{ScreenTime}_{ij} + \beta_{3ij}(\texttt{TimeRdg}_{ij} - \overline{\texttt{TimeRdg}}_j), \tag{1.16}$$

and

$$\begin{aligned}
\text{Level 2:} \qquad \beta_{0j} &= \gamma_{00} + \gamma_{01}\overline{\texttt{TimeRdg}}_j + \gamma_{02}\texttt{Shortages}_j \\
&\quad + \gamma_{03}\texttt{AllFree}_j + U_{0j} \\
\beta_{1j} &= \gamma_{10} \\
\beta_{2j} &= \gamma_{20} \\
\beta_{3j} &= \gamma_{30},
\end{aligned}$$

where $U_{0j} \sim N(0, \tau_{00})$. Although we arrived at this model including weights, we report estimated parameters, standard errors, and various statistics in Table 1.5 for this model with and without weights. Models were also fit to data using only Level 2 weights, and these yielded essentially identical results as those using both Level 1 and Level 2 weights. Only the results with both weights included are reported. This is true here and in later sections. Note that the missing fractions are for the most part small indicating that missing data had a small impact of the results. There are some differences between the results depending on whether weights are used or not. In particular, note that the standard errors are larger when weights are used and the effect of `Girl` is significant at the .05 level when weights are not used but is not significant when weights are used in the estimation.

Before settling on this model and interpreting the parameters, we also preformed some model diagnostics. The U_js were estimated after the model was fit to the data using empirical Bayes estimation. One of the assumptions of the multilevel model was that the U_{pj}s are normal. The estimated U_{0j}s were plotted and found to be roughly normal, which is consistent with the normality assumption but is not conclusive. The normality of estimated U_{0j} only reveals if the assumption is tenable, the assumption may still be violated (Verbeke and Molenberghs, 2000).

We adapted a method, receiver operating characteristic (ROC)[4] curve analysis, used in single

[4]ROC analysis was originally used in signal detection analyses where decisions are made under uncertainty, such as whether a blip on a radar is a signal or noise.

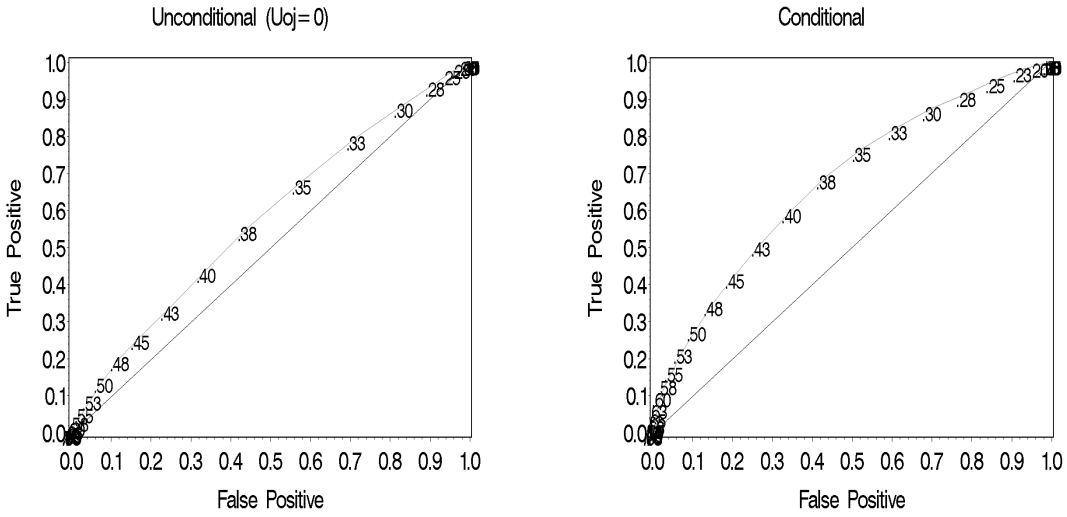

Figure 1.1 *ROC curves for unconditional (i.e., $U_{0j} = 0$) and conditional predictions (i.e., $\hat{U}_{0j}$) based on the final logistic regression model from pseudolikelihood estimation. The numbers are the cut-points used to determine predicted responses from predicted probabilities.*

level logistic regression (Agresti, 2002) to multilevel logistic regression. This analysis gives us a sense of how much is gained from the model and a way to compare models. After a model was fit to data, predicted probabilities setting $U_j = 0$ (i.e., "uncondtional probabilities") were computed and probabilities conditional on the schools (i.e., using estimates of U_{0j}s) were also computed. Predicted responses were determined using multiple cut-points from 0 to 1 incremented by .025. The probabilities of true positives (i.e., $P(\hat{Y}_{ij} = 1|y_{ij} = 1)$) and false positives (i.e., $\hat{Y}_{ij} = 1|y_{ij} = 0)$) were plotted against each other (one point for each cut-point). As an example, the ROC curves for the final model are plotted in Figure 1.1 with the cut-points indicated. The ROC curves for different imputations were nearly indistinguishable so what is plotted in Figure 1.1 is the average of the 15 curves. Chance (random prediction) is indicated by the straight line. The further a model's ROC curve is above the straight line, the better the fit of the model to data. The area under the ROC curves equals the concordance index. Note that conditioning on the schools yields better predictions than not conditioning on the schools, which is evidence for the importance of a random effect for the intercept.

Likelihood ratio tests and information criteria can be used to compare models estimated by maximum likelihood (i.e., $w_{2|j} = w_{1|ij} = 1$). Under MLE, the hypothesis for random effects (e.g., $H_0 : \tau_{00} = 0$ for a random intercept model) is a non-standard test because the conditions for standard test statistics fail (e.g., $\tau_{00} = 0$ is on the boundary of the parameter space). However, variances can be tested by computing a likelihood ratio test statistic but compare it to a mixture of chi-square distributions (for more details see Chapter **??** or Verbeke and Molenberghs, 2000; Molenberghs and Verbeke, 2005; Self and Liang, 1987; Stram and Lee, 1994). Under pseudolikelihood, simple likelihood ratio tests of fixed effects and information criteria for model comparisons should not be used. A correction does exist for likelihood ratio tests (Asparouhov and Muthén, 2006); however, it requires extra computations to obtain the correction factor. The corrected likelihood ratio test is implemented in Mplus and was studied by Asparouhov and Muthén (2006). They found the adjusted likelihood ratio test to be superior to uncorrected one in terms of rejection rates. It should be noted that the corrected test's performance depends on cluster sample size (i.e., larger cluster sizes lead to better results).

Table 1.5 *Estimated parameters, robust standard errors and various statistics for final model fit to binary response variable where weights are not used and both Level 1 and Level 2 weights are used.*

	Maximum likelihood Estimation (no weights)						Pseudolikelihood Estimation (with weights)					
Effect	Estimate	s.e.	Est/s.e.	p	Missing fraction	Odds ratio	Estimate	s.e.	Est/s.e.	p	Missing fraction	Odds ratio
Within or Level 1												
Intercept	−2.617	0.503	5.20	$< .01$	.047	0.07	−3.326	0.684	4.855	$< .01$	.022	0.04
Girl	0.150	0.060	2.50	.01	.042	1.16	0.133	0.071	1.889	.06	.051	1.14
ScreenTime	0.077	0.030	2.56	.01	.084	1.08	0.105	0.038	2.726	.01	.142	1.11
TimeRdg−$\overline{\text{TimeRdg}}$	0.246	0.041	5.95	$< .01$	.041	1.28	0.287	0.057	5.035	$< .01$	.019	1.33
Between or Level 2												
$\overline{\text{TimeRdg}}$	0.860	0.211	5.96	$< .01$	.038	2.36	1.112	0.278	3.998	$< .01$	.018	3.04
Shortages	−0.232	0.081	4.08	$< .01$	.043	0.79	−0.209	0.101	−2.084	.04	.052	.81
AllFree	0.353	0.118	2.86	$< .01$	.047	1.42	0.394	0.153	2.580	.01	.088	1.48
Random Intercept Variance	.250	0.055			.017		.262	0.058	4.483		.021	
	Mean (std dev)											
−2loglikelihood	6698.53 (11.42)											
AIC	6714.53 (11.42)											
BIC	6740.16 (11.42)											

Based on the pseudolikelihood parameter estimates, holding all other variables constant (including the random effect), the odds that a girl will use the internet for school at least weekly are 1.14 time that for boys. In other words, within a given school, girls are more likely than boys to use the Internet to look up information for school. Additionally, students who spend more time in front of a screen (watching TV, DVDs, playing computer or video games) are more likely to use the Internet, and students who spend more time reading for school relative to their classmates are more likely to use the Internet for school.

Student's use of the Internet for school work depends on the school to which they attend. First note that the residual *ICC* for the final model equals $(.262/(0.262+\pi^2/3) =)$ .07. For the sake of comparison when only the Level 1 variables are in the model, $\hat{\tau}_{00} = 0.335$ ($\hat{se} = 0.068$) and the $ICC = 0.335/(0.335+\pi^2/3) = .09$. The between level variables accounted for $(.335 - .262)/.335 \times 100 = 21\%$ of variance of the intercept (differences between schools). Holding observed and unobserved variables constant, the odds ratio that a student will use the Internet for school for a one unit increase in school-mean time spent reading equals 3.04; that is, higher values of $\overline{\text{TimeRdg}}_j$ are associated with larger odds of Internet use. Greater shortages in a school are associated with decreased odds of Internet use; however, larger odds are associated with schools where all students have free or reduced priced lunch.

1.3 Nominal Response Variables

The binary logistic regression model is a special case of the baseline multinomial logistic regression model. After presenting the multinomial model, estimation is briefly discussed followed by an example using the PIRLS data.

1.3.1 The Baseline Multinomial Model

Let k index the K response options such that $Y_{ij} = k$ for $k = 1, \ldots, K$. The Level 1 or cluster-specific model is

$$P(Y_{ij} = k) = \frac{\exp(\beta_{0jk} + \sum_{p=1}^{P} \beta_{pjk} x_{pij})}{\sum_{k=1}^{K} \exp(\beta_{0jk} + \sum_{p=1}^{P} \beta_{pk} x_{pjk})}. \tag{1.17}$$

The sum in the denominator ensures that the sum of the probabilities over response options equals one. Note that there is a different intercept and slope for each response option[5]. This model can become very complex rather quickly. Typically, one response option is chosen as a baseline and the parameters for this response option are set equal to zero. There may be a natural baseline (e.g., "none") or an arbitrary response option can be used as the baseline. We use K here as the baseline or reference category and the β_{pjK}s are set equal to 0.

The Level 2 model is just like the Level 2 model for the binary logistic model and other types of multilevel random effects models, except that there is now a Level 2 model for each Level 1 predictor and each (K-1) response options. The regression coefficients are assumed to be (multivariate) normal; that is,

$$\beta_{0jk} = \sum_{q=0}^{Q} \gamma_{0qk} z_{qj} + U_{0jk} \tag{1.18}$$

$$\vdots$$

$$\beta_{Pjk} = \sum_{q=0}^{Q} \gamma_{Pqk} z_{qj} + U_{Pjk}, \tag{1.19}$$

[5]This model can be re-parameterized as a conditional multinomial logistic regression model such that the predictors can be attributes of *response options*, as well as of individuals (i.e., x_{pijk}). In the conditional model, there is a single β_j for each school and for each response variable x_{ijk}, but there are many more response variables (Agresti, 2002, 2007; Anderson and Rutkowski, 2008).

for $k = 1, \ldots, (K-1)$. The random terms may depend on the response categories as well as the cluster.

The most natural interpretation of the fixed effects parameters is in terms of odds ratios. Given that response K was chosen as the reference category, the cluster-specific model is

$$\log\left(\frac{P(Y_{ij}=k)}{P(Y_{ij}=K)}\right) = \exp\left(\sum_{p=0}^{P}(\sum_{q=0}^{Q}\gamma_{pqk}z_{qj}+U_{pjk})x_{pij}\right). \tag{1.20}$$

The interpretation of main effect and interaction effects are the same as those for the binary logistic regression model; that is, exponentials of γ_{pq}s are estimates of odds ratios.

There are a number of ways to simplify the model. One way is to assume that the random effects do not depend on the category, in which case the k subscript will be dropped (e.g., U_{pj}). Another way to simplify the model is to set some fixed effects to be equal for different response categories (e.g., $\gamma_{pqk} = \gamma_{pqk'}$ where $k \neq k'$). These simplifications are illustrated in Section 1.3.3.

1.3.2 Estimation

Estimation of the multinomial model with weights is basically the same as it was for the binary logistic regression model, except that the binomial distribution is replaced by the more general multinomial distribution and the cluster-specific multilevel nominal logistic regression model replaces the multilevel logistic regression model. With these changes, the cluster-specific log-likelihood is

$$\begin{aligned}\mathcal{L}(y_{ijk}|\boldsymbol{x}_{ij},\boldsymbol{z}_j,\boldsymbol{U}_{jk}) &= \log\left[P(Y_{ij}=1|\boldsymbol{x}_{ij},\boldsymbol{z}_j,\boldsymbol{U}_{jk})^{y_{ij1}}P(Y_{ij}=2|\boldsymbol{x}_{ij},\boldsymbol{z}_j,\boldsymbol{U}_{jk})^{y_{ij2}}\right.\\ &\quad \left.\ldots P(Y_{ij}=K|\boldsymbol{x}_{ij},\boldsymbol{z}_j,\boldsymbol{U}_{jk})^{y_{ijK}}\right],\end{aligned} \tag{1.21}$$

where y_{ijk} is an indicator of the observed response coded such that $y_{ijk} = 1$ if the response from individual i in cluster j is k, and zero otherwise. The cluster-specific multinomial logistic regression model is

$$P(Y_{ij}=k|\boldsymbol{x}_{ij},\boldsymbol{z}_j,\boldsymbol{U}_{jk}) = \frac{\exp\left(\sum_{p=0}^{P}\left(\sum_{q=0}^{Q}\gamma_{pqk}z_{qj}+U_{pjk}\right)x_{pij}\right)}{\sum_{k=1}^{K}\left(\exp\left(\sum_{p=0}^{P}\left(\sum_{q=0}^{Q}\gamma_{pqk}z_{qj}+U_{pjk}\right)x_{pij}\right)\right)}. \tag{1.22}$$

The cluster-specific log pseudolikelihoods are combined as in (1.14) to yield a log pseudolikelihood for the entire data set.

1.3.3 Multinomial Example

The parameters of models reported in this section were fit using SAS/NLMIXED (version 9.3) using the `EMPIRICAL` option to compute the sandwich standard errors. The sandwich standard errors were used in the statistical tests for the fixed effects. In the models, the baseline or reference category was "None or almost never".

Our modeling process started with fitting models with one fixed effect predictor, various combinations of fixed effects, and a simple random structure (i.e., only a single random intercept for each school, $U_{0jk} = U_{0j}$ for all $k = 1, \ldots, 4$). Predictors that were consistently significant were retained in the model. The fixed effects structure settled on is similar to that in the binary model example, except that free lunch was not significant. The results of this random intercept model with complex fixed effects estimated with and without weights are reported in Table 1.6.

Table 1.6 *Estimated parameters, robust standard errors, p value from tests of fixed effects and missing fraction of information for the random intercept multinomial logistic regression model without weights (MLE) and with Level 1 and 2 weights (pseudolikelihood).*

	Maximum Likelihood Estimation (no weights)				Pseudolikelihood Estimation (with weights)			
Effect	Est	s.e.	p	Missing fraction	Est	s.e.	p	Missing fraction
Intercept 1	−3.949	0.720	< .001	.045	−5.320	1.005	< .001	.024
Intercept 2	−3.169	0.729	< .001	.024	−3.673	1.064	< .001	.012
Intercept 3	−1.714	0.696	.014	.022	−2.195	0.986	.026	.007
Girl 1	0.411	0.101	< .001	.029	0.387	0.120	.001	.046
Girl 2	0.332	0.080	< .001	.032	0.309	0.100	.002	.037
Girl 3	0.453	0.072	< .001	.018	0.422	0.096	< .001	.019
Screen Time 1	0.161	0.051	.002	.207	0.165	0.066	.011	.232
Screen Time 2	0.006	0.039	.868	.084	0.026	0.047	.575	.164
Screen Time 3	−0.047	0.037	.200	.163	−0.068	0.039	.084	.167
CTime Rdg 1	0.377	0.069	< .001	.041	0.444	0.094	< .001	.029
CTime Rdg 2	0.251	0.055	< .001	.031	0.275	0.068	< .001	.017
CTime Rdg 3	0.129	0.055	.020	.044	0.129	0.064	.043	.026
MTime Rdg 1	1.182	0.300	< .001	.029	1.717	0.409	< .001	.015
MTime Rdg 2	1.274	0.302	< .001	.022	1.471	0.435	< .001	.012
MTime Rdg 3	0.704	0.288	.015	.020	0.938	0.405	.020	.008
Shortages 1	−0.229	0.138	.098	.020	−0.177	0.141	.211	.035
Shortages 2	−0.364	0.111	.001	.033	−0.378	0.142	.007	.041
Shortages 3	−0.267	0.134	.047	.023	−0.275	0.149	.063	.034
Variance	0.315	0.063	—	.023	0.325	0.065	—	.057
	Mean (Std Dev)							
−2lnlike	13,565.27 (11.20)							
AIC	13,603.27 (11.20)							
BIC	13,664.15 (11.20)							

Possible simplifications of the model are suggested by examining the parameters reported in Table 1.6. For example, the parameter estimates for `Girl` are similar in value for the first three response options, and only the parameters for the first response option for `ScreenTime` is significant. Parameters that are similar in value can be equated and those that are not significant can be set to 0 (i.e., equated to the value for the baseline category). Significance tests can be computed for these restrictions using Wald tests. We can test whether the coefficients for effects are all the same, such as those for `Girl`, $H_0 : \gamma_{11} = \gamma_{12} = \gamma_{13}$. Let P^* equal the number of fixed effects parameters and $\boldsymbol{L}$ equal an $(r \times P^*)$ matrix whose rows define linear combinations of the P^* parameters, $\boldsymbol{\gamma}$ is a $(P^* \times 1)$ vector of parameters, and $\boldsymbol{S}_{\hat{\gamma}}$ the covariance matrix of parameter estimates. The null hypothesis for `Girls` is

$$H_0 : \boldsymbol{L\gamma} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & -1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 & \dots & 0 \end{pmatrix} \begin{pmatrix} \gamma_{00} \\ \gamma_{01} \\ \gamma_{02} \\ \gamma_{03} \\ \gamma_{11} \\ \gamma_{12} \\ \gamma_{13} \\ \vdots \\ \gamma_{P3} \end{pmatrix} = \begin{pmatrix} \gamma_{11} - \gamma_{12} \\ \gamma_{11} - \gamma_{13} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

and

$$\text{Wald} = \boldsymbol{\gamma' L'} (\boldsymbol{L S_{\hat{\gamma}} L'})^{-1} \boldsymbol{L\gamma} \sim \chi^2_r.$$

For the above hypothesis and using the sandwich covariance matrix for $\boldsymbol{S}_{\hat{\gamma}}$, Wald = 1.34 with degrees of freedom $r = 2$ and $p = .51$, which supports the conclusion that a single coefficient for `Girl` is sufficient. Testing whether the coefficients for `CTimeRdg` are the same yields Wald = 13.98, $df = 2$ and $p < .01$ and indicates that these should not be equated. Note that when testing a single fixed effect, such as $H_o : \gamma_{pq} = 0$, the test statistic reduces to $(\hat{\gamma}_{pq} / \widehat{se}_{\gamma_{pq}})^2$.

Based on a number of Wald tests, in the next round of modeling, some parameters were equated and others were set equal to 0. Given a simpler fixed effects structure, the random structure of the model was developed. Category-specific random effects were first added to the model that only had an intercept. The variances of U_{0j1} and U_{0j2} were similar in value and the estimated correlation between them equaled .79. This suggested setting $U_{0j1} = U_{0j2}$ and implies that the random effect for a school is the same when the response is either "Every day or almost every day" and "Once or twice a week". With this restriction, the model with both the fixed effects structure found previously and a semi-complex random structure was fit to the data and selected as our final model. The parameter estimates and various statistics for the final model are reported in Table 1.7. The models fit by MLE reported in Tables 1.6 and 1.7 can be compared using AIC and BIC. The model in Table 1.7 fit by MLE with the more complex random structure is better in terms of AIC and BIC than they are in Table 1.6.

Table 1.7 *Estimated parameters, robust standard errors, p value from tests of fixed effects, missing fraction of information, and estimated odds ratios for the final multinomial logistic regression model fit using MLE (no weights) and pseudolikelihood estimation (with weights).*

	Maximum Likelihood Estimation (no weights)					Pseudolikelihood Estimation (with weights)				
Effect	Est	s.e.	p	Missing fraction	Odds ratio	Est	s.e.	p	Missing fraction	Odds ratio
Intercept 1	−4.259	0.662	< .001	.031	0.01	−5.274	0.989	< .001	.018	0.01
Intercept 2	−3.252	0.656	< .001	.021	0.04	−4.154	0.966	< .001	.012	0.02
Intercept 3	−1.824	0.670	.007	.019	0.16	−2.268	0.945	.016	.009	0.10
Girl 1	0.409	0.062	< .001	.026	1.52	0.383	0.080	< .001	.029	1.47
Girl 2	0.409	.			1.52	0.383	.			1.47
Girl 3	0.409	.			1.52	0.383	.			1.47
Screen Time 1	0.166	0.044	< .001	.218	1.18	0.170	0.056	.003	.237	1.19
Screen Time 2	0.000	.			1.00	0.000	.			1.00
Screen Time 3	0.000	.			1.00	0.000	.			1.00
CTime Rdg 1	0.382	0.069	< .001	.042	1.47	0.452	0.095	< .001	.031	1.57
CTime Rdg 2	0.256	0.055	< .001	.032	1.29	0.281	0.070	< .001	.019	1.32
CTime Rdg 3	0.127	0.055	.021	.046	1.14	0.126	0.063	.046	.026	1.13
MTime Rdg 1	1.242	0.279	< .001	.022	3.46	1.632	0.405	< .001	.013	5.12
MTime Rdg 2	1.242	.			3.46	1.632	.			5.12
MTime Rdg 3	0.679	0.285	.017	.018	1.97	0.893	0.395	.024	.008	2.44
Shortages 1	0.000	.			1.00	0.000	.			1.00
Shortages 2	−0.180	0.084	.033	.038	0.84	−0.234	0.108	.031	.061	0.79
Shortages 3	−0.180	.			0.84	−0.234	.			0.79
$\tau_{11} = \tau_{22}$	0.436	0.081		.030		0.456	0.085		.053	
τ_{33}	0.315	0.070		.011		0.304	0.075		.029	
$\tau_{13} = \tau_{23}$	0.246	0.065	< .001	.023		0.255	0.068		.060	
	Mean (Std Dev)									
−2lnlike	13,515.02 (12.42)									
AIC	13,543.10 (12.42)									
BIC	13,587.87 (12.42)									

Even after simplifying the model, there are still a large number of fixed effects parameters. To aid interpretation and in some circumstances. it may be desirable to compare all pairs of response options. All possible odds ratios can be computed by calculating for all unique pairs of response options. For example, suppose that we wish to look at the effect of school-mean centered reading time on the odds of "Every day or almost every day" (Daily) versus "Once to twice a week" (Weekly). Based on the pseudolikelihood estimates, this odds ratio equals $\exp(0.452 - 0.281) = 1.19$. Since there were no interactions in the final model, all the odds ratios were computed using $\exp(\gamma_{pqk} - \gamma_{pqk'})$ and are reported in Table 1.8. For each effect, the odds ratios below the diagonals were computed using the pseudolikelihood estimates and those above were from MLE. An entry in the table is the odds ratio of more frequent use of the Internet versus less frequent use for unit change of the predictor variable. For example, the pseudolikelihood estimate of the odds ratio of Daily use of the Internet versus Monthly for a one unit change in school-mean centered reading time is 1.39 and the maximum likelihood estimate is 1.29.

Not all the pairwise odds ratios are necessarily significant. A test for an odds ratio can be conducted or a confidence interval can be placed on the odds ratios. In general, a test statistic for $H_o : \exp(\gamma_{pqk}) = \exp(\gamma_{pqk'})$ equals

$$\frac{\hat{\gamma}_{pqk} - \hat{\gamma}_{pqk'}}{\hat{se}_{(\hat{\gamma}_{pqk} - \hat{\gamma}_{pqk'})}},$$

where that standard error of the difference equals

$$\widehat{se}_{(\hat{\gamma}_{pqk} - \hat{\gamma}_{pqk'})} = \sqrt{\text{var}(\hat{\gamma}_{pqk}) + \text{var}(\hat{\gamma}_{pqk'}) - 2\text{cov}(\hat{\gamma}_{pqk}, \hat{\gamma}_{pqk'})},$$

and sandwich estimates are used for the variances and covariances. Given the large sample sizes, the above test statistic can be compared to a standard normal distribution. In Table 1.8, the significant odds ratios are in bold face. For a $(1 - \alpha) \times 100\%$ confidence interval for an odds ratio, an interval is first computed for the γs,

$$(\hat{\gamma}_{pk} - \hat{\gamma}_{pk'}) \pm z_{\alpha/2} \widehat{se}_{(\hat{\gamma}_{pk} - \hat{\gamma}_{pk'})},$$

and then the exponential of the endpoints taken for the interval of the odds ratio.

The ordinal nature of the responses is apparent in Table 1.8. In general, the odds are non-decreasing as the reported amount of time using the Internet decreases (i.e., odds ratios tend to increase as one goes down the columns for the pseudolikelihood or across the rows for the maximum likelihood ones). The exception is for $\texttt{shortages}_j$ where the odds ratios are greater than 1 for Daily versus Weekly or Monthly, but they are less than 1 for Weekly or Monthly versus Never. The direction of the effect of $\texttt{shortages}_j$ is different for different pairs of response options. This illustrates one of the strengths of the multnomial model. The multinomial model permits a fine grained analysis, including the possibility of reversals in the direction of effects.

1.4 Ordinal Response Variables

When response options have a natural ordering, as they do in our PIRLs data, the ordering can be incorporated into an analysis by selecting a model that explicitly uses the ordering of the response options. In ordinal models, the response options are dichotomized based on the ordering of the categories, which yields $(K - 1)$ dichotomies. The major difference between ordinal models is how responses are dichotomized and whether restrictions are placed on parameters over the dichotomies. Three of the most common ordinal models are continuation ratios, adjacent categories and proportional odds models. These are presented in Sections 1.4.1, 1.4.2 and 1.4.3. Less restrictive versions of the adjacent categories and proportional odds models are also discussed. Throughout this section is it assumed that categories are ordered from $k = 1$ to K.

Table 1.8 *All possible odds ratios from final model of more frequent versus less frequent where those below the diagonals are estimated using pseudolikelihood (with weights) and those above the diagonal from maximum likelihood estimation (no weights). The odds ratios in bold are significantly different from* 1 *and those in boxes are adjacent categories.*

Predictor Variable	Response Options	Response Options: Daily	Weekly	Monthly	Never
Girl	Daily		1.00	1.00	**1.52**
	Weekly	1.00		1.00	**1.52**
	Monthly	1.00	1.00		**1.52**
	Never	**1.47**	**1.47**	**1.47**	
Screen Time	Daily		**1.18**	**1.18**	**1.18**
	Weekly	**1.19**		1.00	1.00
	Monthly	**1.19**	1.00		1.00
	Never	**1.19**	1.00	1.00	
Centered Time Reading	Daily		**1.35**	**1.29**	**1.47**
	Weekly	**1.19**		**1.14**	**1.29**
	Monthly	**1.39**	**1.17**		**1.14**
	Never	**1.57**	**1.32**	**1.13**	
Mean Time Reading	Daily		1.00	**1.75**	**3.46**
	Weekly	**1.00**		**1.75**	**3.46**
	Monthly	**2.09**	**2.10**		**1.97**
	Never	**5.11**	**5.12**	**2.44**	
Shortages	Daily		**1.20**	**1.20**	1.00
	Weekly	**1.26**		1.00	**0.84**
	Monthly	**1.26**	1.00		0.84
	Never	1.00	**0.79**	**0.79**	

1.4.1 Continuation Ratios

One common and simple approach for ordinal responses is to form continuation ratios,

$$\frac{P(Y_{ij}=k)}{P(Y_{ij}=k+1)+\ldots+P(Y_{ij}=K)} \quad \text{for} \quad k=1,\ldots,K-1, \tag{1.23}$$

or

$$\frac{P(Y_{ij}=k+1)}{P(Y_{ij}=1)+\ldots+P(Y_{ij}=k)} \quad \text{for} \quad k=1,\ldots,K-1. \tag{1.24}$$

With continuation ratios, multilevel binary logistic regression models are fit separately to each of the $(K-1)$ ratios.

For the PIRLS data, multilevel logistic regression models could be fit to each probability of a more frequent use of the Internet versus a less frequent use; that is, multilevel binary logistic regression models could be fit to each of the following ratios:

$$\text{Ratio I} \qquad P(\text{Daily})/P(\text{Weekly, Monthly or Never}) \tag{1.25}$$

$$\text{Ratio II} \qquad P(\text{Weekly})/P(\text{Monthly or Never}) \tag{1.26}$$

$$\text{Ratio III} \qquad P(\text{Monthly})/P(\text{Never}). \tag{1.27}$$

Advantages of this approach over the other ordinal models discussed in this section is that binary

logistic regression models are much simpler to use and different models may be found for each way of forming odds ratios. A relative disadvantage of this model is that except for one ratio (e.g., (1.25)), the models for the other ratios (e.g., (1.26) and (1.27)) are based on a sub-set of the data and hence smaller sample sizes.

1.4.2 Adjacent Categories

A model that compares pairs of response options is the adjacent categories logit model (Hartzel et al., 2001). The cluster-specific (Level 1) model of the multilevel adjacent categories logit models is

$$\log\left(\frac{P(Y_{ij}=k)}{P(Y_{ij}=k+1)}\right)=\sum_{p=0}^{P}\beta_{pjk}x_{pij} \quad \text{for} \quad k=1,...,K-1. \tag{1.28}$$

It may be reasonable that the effect of predictors are the same for each pair of adjacent responses; therefore, in the Level 2 models, the fixed effects are specified so that they do not depend on k. Only the fixed effects for the Level 2 intercept of the Level 1 intercept are allowed to depend on k. The Level 2 model for the intercept is

$$\beta_{0jk}=\gamma_{00k}+\sum_{q=1}^{Q}\gamma_{0q}z_{qj}+U_{0jk} \quad \text{for} \quad k=1,\ldots K-1. \tag{1.29}$$

From (1.29), we can see that the intercept of the intercepts γ_{00k} and the random effects U_{0jk} can differ over pairs of adjacent logits, but the coefficients for the predictors of the intercept γ_{0q} are fixed over k.

The Level 2 models for the Level 1 predictors x_{pij} for $p>0$ are

$$\beta_{pjk}=\sum_{q=0}^{Q}\gamma_{pq}z_{qj}+U_{pjk} \quad \text{for} \quad k=1,\ldots K-1. \tag{1.30}$$

Note that the fixed effects do not depend on k, but the random effects may depend on the specific pair of responses, k and $k+1$.

To show that the fixed effects of the predictors are the same for neighboring categories, we replace the β_{pjk}s in the Level 1 model (1.28) by their Level 2 models, (1.29) and (1.30), and obtain

$$\log\left(\frac{P(Y_{ij}=k)}{P(Y_{ij}=k+1)}\right)=\underbrace{\gamma_{00k}+\sum_{p=0}^{P}U_{pjk}}_{\text{Depends on } k}+\underbrace{\sum_{q=1}^{Q}\gamma_{0q}z_{qj}+\sum_{p=1}^{P}\sum_{q=0}^{Q}\gamma_{pq}z_{qj}x_{pij}}_{\text{Does not depend on } k}. \tag{1.31}$$

The fixed effects for x_{pij} and z_{qj} are the same for all adjacent categories. The exponentials of the γ_{pq}s are interpretable as odds ratios just as they are in binary and multinomial logistic regression. The restriction that the fixed effects of x_{pij} and z_{qj} are the same for neighboring responses is a strong assumption.

To estimate the adjacent categories model, we use the fact that the model is a special case of the baseline multinomial model. For the adjacent categories model, the log of the likelihood and the model for the $P(Y_{ij}=k|\boldsymbol{x}_{ij},\boldsymbol{z}_j,\boldsymbol{U}_{jk})$ are the same as those for the baseline multinomial model; however, linear restrictions must be placed on the fixed effects parameters of the multinomial model. To see the connection between these models and to find the proper restrictions on the multinomial parameters, we first note that the baseline logits can be written as the sum of the adjacent category logits; that is,

$$\log\left(\frac{P(Y_{ij}=k)}{P(Y_{ij}=K)}\right)=\log\left(\frac{P(Y_{ij}=k)}{P(Y_{ij}=k+1)}\right)+\ldots+\log\left(\frac{P(Y_{ij}=K-1)}{P(Y_{ij}=K)}\right), \tag{1.32}$$

for $k = 1, \ldots, K-1$. Substituting the adjacent categories model (1.31) into (1.32) and simplifying yields

$$\log\left(\frac{P(Y_{ij}=k)}{P(Y_{ij}=K)}\right) = \sum_{h=k}^{K-1}\gamma_{00h} + \sum_{p=0}^{P}\left(\sum_{h=k}^{K-1}U_{pjh}\right)x_{pij} + (K-k)\left(\sum_{q=1}^{Q}\gamma_{0q}z_{qj} + \sum_{p=1}^{P}\sum_{q=0}^{Q}\gamma_{pq}z_{qj}x_{pij}\right). \tag{1.33}$$

Equation (1.33) is a restricted version of the multinomial logistic regression model. The correspondence between parameters of the two models is as follows where the multinomial parameters have the superscript "[mlt]" and the adjacent categories parameters have no superscript:

Multinomial Model		Adjacent Categories
$\gamma_{00k}^{[mlt]}$	=	$\sum_{h=k}^{K-1}\gamma_{00h}$
$\gamma_{pqk}^{[mlt]}$	=	$(K-k)\gamma_{pq}$ for $p > 01$
$U_{pjk}^{[mlt]}$	=	$\sum_{h=k}^{K-1}U_{pjh}$.

Therefore, the adjacent categories intercept parameters equal $\gamma_{00k} = \gamma_{00k}^{[mlt]} - \gamma_{00,k-1}^{[mlt]}$. The fixed effects for predictors in the adjacent categories model equal $\gamma_{pq} = \gamma_{pqk}^{[mlt]} - \gamma_{pq,k-1}^{[mlt]}$; that is, the difference $\gamma_{pqk}^{[mlt]} - \gamma_{pq,k-1}^{[mlt]}$ must be restricted to equal a constant value for a given p and q. This restriction implies that the odds ratios for adjacent categories are the same regardless of which two neighboring categories are compared (i.e., $\exp[(\gamma_{pq,k+1}^{[mlt]} - \gamma_{pqk}^{[mlt]})x_{pij}] = \exp[\gamma_{pq}x_{pij}]$). The odds ratios do not depend on which two adjacent categories are compared, which is a property of the adjacent categories model. It follows that if a multinomial model has been fit to data, the results can provide information about the plausibility of an adjacent categories model, in particular, whether the effects of the x_{pij}s and z_{qj}s depend on the response options.

For the PIRLS example on Internet usage, we can use our results from the baseline multinomial model described in Section 1.3 and examine the estimated odds ratios for adjacent categories that are in boxes in Table 1.8. If the adjacent categories logit models holds, then the odds ratios for different response options for a predictor should be equal. The only odds ratios in Table 1.8 that are comparable in value are those for school-mean centered time reading (i.e., for pseudo-likelihood estimation, 1.19, 1.17 and 1.13, and for MLE, 1.35, 1.14 and 1.14). The fact that some of the adjacent categories odds ratios for predictors are significantly different from 1 but others for the same predictor equal 1 implies that the adjacent categories model is not appropriate. For example, the odds ratio for `Girls` comparing Daily versus Weekly and Weekly versus Monthly both equal 1, but Monthly versus Never equals 1.47, which is significantly different from 1. Since the adjacent categories models parameters equal multinomial model parameters with linear restrictions on them, Wald tests as described in Section 1.3.3 can be constructed to test the restrictions implied by the adjacent categories models. However, given the odds ratios in Table 1.8, it is unlikely that the assumption of equal effects of the predictors will hold, except for school-mean centered time reading. This implies the adjacent category model will not fit the data well. Although we go no further with this example, one possible model that would be interesting to investigate is a "*partial adjacent categories*" model where some but not all the predictors have equal effects for adjacent response options.

1.4.3 Cumulative Probabilities

A third common choice for ordinal response data is the proportional odds model where the cumulative probabilities, $P(Y_{ij} \leq k)$, are modeled using a multilevel logit formulation. The cluster specific

(Level 1) proportional odds model is

$$\log\left(\frac{P(Y_{ij}\leq k)}{P(Y_{ij}>k)}\right)=\beta_{0jk}+\sum_{p=1}^{P}\beta_{pj}x_{ij}\quad\text{for}\quad k=1,\ldots,K-1, \tag{1.34}$$

where the intercepts have the same order as the response options (i.e., $\beta_{0j1}\leq\beta_{0j2}\leq\ldots\leq\beta_{0jK}$), and the cluster-specific regression coefficients for the predictors are the same regardless of the response option.

Similar to the adjacent categories model, since there is a single regression coefficient for each x_{pij}, the effect of x_{pij} is the same regardless of which cumulative odds ratio is examined. In other words, the odds ratios for a unit change in x_{pij} (where x_{pij} only has a main effect) equals

$$\frac{P(Y_{ij}\leq k|\boldsymbol{x}_{-p,ij},x_{pij})}{P(Y_{ij}>k|\boldsymbol{x}_{-p,ij},x^{*}_{pij})}=\exp[\beta_{pj}(x_{pij}-x^{*}_{pij})], \tag{1.35}$$

where x_{pij} and x^{*}_{pij} are two values of predictor p, and $\boldsymbol{x}_{-p,ij}$ are the remaining predictors that are held fixed to some value. For a one unit change in a predictor (i.e., $(x_{pij}-x^{*}_{pij})=1$), the odds ratios equal $\exp(\beta_{pj})$. The odds ratios do not depend on the response option and they only depend on the difference between two values of the predictor variable and the value of β_{pj}.

The Level 2 models for the cluster-specific regression coefficients for β_{pj} are the same as those from the adjacent categories model (i.e., (**??**) and (1.30)), except that the random effects do not depend on k. Only the Level 2 intercept of the Level 1 intercept depends on the response option k (i.e., γ_{00k}). Furthermore, the order of these fixed effects γ_{00k} reflect the ordering of the response options; that is, $\gamma_{001}\leq\gamma_{002}\leq\ldots\leq\gamma_{00K}$.

The proportional odds model also has a latent variable interpretation. As in Section 1.2.2, a latent variable Y^{*}_{ij} is proposed; however, since there are now multiple categories of the response variable, there are multiple ordered thresholds that determine the observed response based on the latent variable. Specifically,

$$Y_{ij}=\begin{cases}1 & Y^{*}_{ij}\leq\xi_{j1}\\ 2 & \xi_{j1}<Y^{*}_{ij}\leq\xi_{j2}\\ \vdots & \\ K & \xi_{j,K-1}<Y^{*}_{ij},\end{cases}$$

and

$$Y^{*}_{ij}=\sum_{q=1}^{Q}\gamma_{0q}z_{qj}+\sum_{p=1}^{P}(\sum_{q=0}^{Q}\gamma_{pq}z_{j}+U_{pj})x_{pij}+\varepsilon_{ij}+U_{0j}$$

The above model for $Y*_{ij}$ does not have a fixed intercept γ_{00k}. The thresholds equal the negative of the fixed intercept γ_{00k} in the Level 2 model for β_{00k}; that is, $\gamma_{00k}=-\xi_{jk}$. In this latent variable formulation, Y^{*}_{ij} represents an individual's value along some underlying continuum. The individual's value may randomly due to ε_{ij} and U_{j}, but the thresholds are fixed and ordered. The distribution of ε_{ij} determines the probability model for the observed responses (i.e., normal or logistic)[6].

Estimation and the procedure to incorporated weights in the estimation of the proportional odds models is the same as that for the baseline multinomial logistic regression models described in Section 1.3.2. The only difference is that the probabilities $P(Y_{ij}=1|\boldsymbol{x}_{ij},\boldsymbol{z}_{j},\boldsymbol{U}_{j})$ based on the cluster-specific proportional odds model are used in (1.21) rather than those based on the multinomial model. To represent the probabilities, we first note that replacing the β_{pj}s in the Level 1 model by their Level 2 models gives us the cluster-specific model for the cumulative probabilities,

$$\log\left(\frac{P(Y_{ij}\leq k|\boldsymbol{x}_{ij},\boldsymbol{z}_{j},\boldsymbol{U}_{j})}{P(Y_{ij}>k|\boldsymbol{x}_{ij},\boldsymbol{z}_{j},\boldsymbol{U}_{j})}\right)=\gamma_{00k}+\sum_{q=0}^{Q}\gamma_{0q}z_{qj}+U_{0j}+\sum_{p=1}^{P}\sum_{q=0}^{Q}(\gamma_{pq}+U_{qj})x_{ij}\quad\text{for}\quad k=1,\ldots,K-1.$$

[6] A random utility model can also be proposed where the latent variables also depend on the response option and a person chooses the response that have the largest value Y^{*}_{ijk}. See the footnote in Section 1.2.2 for more details.

Table 1.9 *Parameter estimates of proportional odds models with and without weights. The standard errors are empirical (sandwich) ones.*

	Maximum Likelihood (no weights)					Pseudolikelihood (weights)				
Effect	Est	s.e.	p	Odds Ratio	Missing Fraction	Est	s.e.	p	Odds Ratio	Missing fraction
Intercept 1	−4.179	0.492	< .01	0.01	.03	−5.108	0.738	< .01	0.02	.02
Intercept 2	−2.791	0.488	< .01	0.03	.04	−3.681	0.733	< .01	0.03	.02
Intercept 3	−1.573	0.484	< .01	0.09	.03	−2.432	0.731	< .01	0.09	.01
Girl	0.252	0.055	< .01	1.26	.04	0.233	0.064	< .01	1.26	.05
ScreenT 1	0.071	0.028	.01	1.08	.15	0.080	0.036	.03	1.08	.20
CTimeRdg 1	0.257	0.042	< .01	1.34	.03	0.296	0.056	< .01	1.34	.01
MTimeRdg	0.927	0.203	< .01	3.58	.03	1.275	0.300	< .01	3.58	.01
Shortages 1	−0.209	0.086	.02	0.83	.03	−0.182	0.095	.06	0.83	.04
Variance	0.240	0.049			.02	0.251	0.049			.03
	Mean (Std Deiviation)									
−2loglike	13608.41 (11.43)									
AIC	13626.41 (11.43)									
BIC	13626.44 (11.43)									

Using these cluster-specific cumulative probabilities, the probability of a specific response option is found by subtraction as follows:

$$P(Y_{ij} = k|\boldsymbol{x}_{ij}, \boldsymbol{z}_j, \boldsymbol{U}_j) = P(Y_{ij} \leq k|\boldsymbol{x}_{ij}, \boldsymbol{z}_j, \boldsymbol{U}_j) - P(Y_{ij} \leq (k-1)|\boldsymbol{x}_{ij}, \boldsymbol{z}_j, \boldsymbol{U}_j),$$

and $P(Y_{ij} = 1|\boldsymbol{x}_{ij}, \boldsymbol{z}_j, \boldsymbol{U}_j) = P(Y_{ij} \leq 1|\boldsymbol{x}_{ij}, \boldsymbol{z}_j, \boldsymbol{U}_j)$.

1.4.4 Example

For the proportional odds model, we used the same predictors as used for the baseline multinomial model. The cumulative probabilities of more frequent to less frequent were modeled; that is,

$$\begin{aligned} P(Y_{ij} = \text{Daily})/P(Y_{ij} = \text{Daily, Weekly, Monthly or Never})) \\ P(Y_{ij} = \text{Daily or Weekly})/P(Y_{ij} = \text{Monthly or Never}) \\ P(Y_{ij} = \text{Daily, Weekly or Monthly})/P(Y_{ij} = \text{Never}) \end{aligned}$$

The model fit to the data was

$$\begin{aligned} \log\left(\frac{P(Y_{ij} \leq k)}{P(Y_{ij} > k)}\right) &= \gamma_{0k} + \gamma_{10}\texttt{Girl}_{ij} + \gamma_{20}\texttt{ScreenTime}_{ij} + \gamma_{30}\texttt{CTimeRdg}_{ij} + \gamma_{40}\overline{\texttt{MTimeRdg}}_j \\ &\quad + \gamma_{50}\texttt{Shortages}_j + U_{0j}, \end{aligned}$$

where $U_{0j} \sim N(0, \tau_{00})$ *i.i.d*.

The estimated parameters are reported in Table 1.9. Unlike the baseline multinomial model, the predictor $\texttt{ScreenTime}_{ij}$ is no longer significant and whether design weights are incorporated or not leads to different conclusions for $\texttt{Shortages}_j$ (i.e., it is not significant for pseudolikelihood but is significant for maximum likelihood). One possibility is the effect of predictors are not the same over response options. Proportional odds is a strong assumption. A *partial proportional odds model* can be fit by relaxing the equal slopes assumption for some or all of the predictors (Peterson and Harrell, 1990); that is, allow β_{ij} to depend on the response options. The equality of the predictors can be tested in the same way as described in Section 1.3.3. For the PIRLS example, response option dependent γ_{pqk}s were estimated for $\texttt{ScreenTime}_{ij}$, $\texttt{CTimeRdg}_{ij}$, and $\texttt{Shortages}+j$. Models

Table 1.10 *Parameter estimates of partial proportional odds models with weights (pseudolikelihood) where standard errors are empirical (sandwich) ones. A "." for the standard error indicates that this parameter was fixed or equated with another.*

Effect	Est	s.e.	p	Est	s.e.	p	Odds Ratio	Fraction missing
Intercept 1	−5.512	0.772	< .01	−5.378	0.747	< .01	0.01	0.01
Intercept 2	−3.783	0.752	< .01	−3.801	0.744	< .01	0.02	0.01
Intercept 3	−2.133	0.754	< .01	−2.219	0.732	< .01	0.11	0.01
Girl	0.231	0.065	< .01	0.232	0.064	< .01	1.26	0.04
ScreenT 1	0.196	0.059	< .01	0.145	0.031	< .01	1.16	0.14
ScreenT 2	0.127	0.038	< .01	0.145	.		1.16	
ScreenT 3	−0.013	0.039	.73	0.000	.		1.00	
CTimeRdg 1	0.338	0.086	< .01	0.298	0.056	< .01	1.35	0.02
CTimeRdg 2	0.311	0.059	< .01	0.298	.		1.35	
CTimeRdg 3	0.263	0.061	< .01	0.298	.		1.35	
MTimeRdg	1.266	0.308	< .01	1.274	0.306	< .01	3.58	0.01
Shortages 1	0.042	0.122	.73	0.000	.		1.00	
Shortages 2	−0.167	0.099	.09	−0.227	0.074	< .01	0.80	0.06
Shortages 3	−0.286	0.117	.01	−0.227	.		0.79	
Variance	0.263	0.051		0.255	0.050			0.04

with different coefficients for $\texttt{MTimeRdg}_j$ failed to converge. The results from the proportional odds model that relaxes the assumption for three predictors are reported for pseudolikelihood estimation on the left side of Table 1.10. The equality of γ_{pqk}s were tested and the proportional odds assumption appeared to be valid for only school-mean centered reading time. The tests also indicated that some of the slopes for other effects could be combined and others could be set equal to 0. A final model was fit to the data incorporating the changes suggested by the tests. When estimated by maximum likelihood estimation, this final partial proportional odds model has the smallest BIC among the models fit to cumulative odds and has essentially the same AIC as the other model in Table 1.10, which are both smaller than the AIC from the proportional odds model. The results for the final model are on the right-side of Table 1.10.

Over all the analysis on this data set, the basic story is the same. Holding other predictors constant, the odds of more regular usage of the Internet for school is higher for girls than boys, larger for students who spend more time using electronic entertainment (`ScreenTime`), larger for students who read more for homework, and larger for students in schools where the average time spend by students reading for homework is larger. The effect of shortages appears somewhat mixed. The parameter estimates from the binary logistic model and the proportional odds suggest that an increase in shortages is associated with a decrease in the odds of more regular usage of the Internet. The partial proportional odds model indicates that the odds ratio for shortages equals 1 when comparing Daily usage versus less regular usage; however, the other two cumulative odds ratio equal 0.79, which is more similar to the binary and proportional odds models. The results from the baseline multinomial model help to explain the conflicting results. The estimated odds ratios from the multinomial model in Table 1.8 for Daily versus Weekly and for Daily versus Monthly both equal 1.26 (i.e., an increase in shortages is associated with an increase in odds); however, the odds ratios for Weekly versus Never and for Monthly versus Never both equal 0.79 (i.e, an increase in shortages is associated with a decrease in the odds of usage of the Internet). Lastly, the odds ratio for Weekly versus Monthly equals 1. The direction of the effect of shortages changes direction depending on which response options are compared. How to form logits and which model should be used depends on the researcher's goal, hypotheses, and how the results will be used.

1.5 Software

Any section on software will quickly become out of date; however, we briefly mention those programs that are currently available and meet the following criteria: it is capable of estimating models parameters using adaptive quadrature, can easily compute sandwich standard errors, and it can incorporate weights. Based on the current state of the art (and our knowledge), this list includes SAS/NLMIXED (SAS Institue Inc, 2011a), STATA/GLAMM (Rabe-Hesketh and Skronkal, 2012), and Mplus (Muthén and Muthén, 1998-2010).

Mplus is capable of estimating parameters for the random effects binary logistic regression and proportional odds models, but it not capable of estimating random effects multinomial, adjacent categories, partial proportional odds models, or variations of these models. SAS/NLMIXED and STATA/GLAMM can estimate a wider range of models. The procedure NLMIXED afforded a great deal of flexibility with minimal programing effort, including restrictions that could be placed on model parameters. Grilli and Pratesi (2004) described how to use SAS/NLMIXED for the binary logistic regression and proportional odds models; however, we modified and simplified it for the models discussed in this chapter. Grilli and Pratesi (2004) used more complicated procedures to retain multiple decimal places of the Level 2 weights and to obtain reasonable standard errors. This complexity is not necessary. At least for the current release of SAS (version 9.3), the incorporation of Level 2 weights does not require any tricks because the command to include the Level 2 weights permits non-integer values[7]. The second difficulty that Grilli and Pratesi (2004) encountered was the computation of standard errors; however, sandwich estimates may be easily obtained by using the NLMIXED option EMPIRICAL. STATA/GLAMM can fit models with more than two levels; whereas, Mplus and SAS/NLMIXED can only deal with two level models.

1.6 Discussion

A wealth of information is available from large scale national and international survey data, much of which is open source. Although many surveys are designed to measure educational achievement, a large number of surveys and items on even those created to measure educational attainment can be used to study a variety of topics. For example, in this chapter we studied Internet usage for school work by fourth grade students. Given that questions often have categorical response options, they should be modeled using a model designed for discrete data. Which specific model a researcher uses depends on what is appropriate for the data and meets the researcher's goal and hypotheses. For example, if a researcher wants to contrast or compare response options for pairs of ordered categories, then the adjacent categories model would be useful; however, if one wants to make statements about effect above or below various points on the response scale, a (partial) proportional odds model might be the best choice. Substantive considerations are of paramount importance, but so is taking into consideration the nature and characteristics of the data at hand.

The methodology for analyzing complex survey data with multilevel models for discrete response variables is a quickly changing field. This chapter presented what we feel is the current state of affairs. Additional software options for fitting multilevel survey data with design weights are likely to become available over time. Furthermore, the methods developed for missing data in the context of multilevel models is an active area of research and we expect that more efficient methods than what was employed here will become available (e.g., Swoboda, 2011; Kim and Swoboda, 2012, April). Regardless of these shortcomings, models, methodology and tools exist to analyze discrete response data from large scale survey data in an appropriate manner.

[7]The SAS/NLMIXED documentation describing the command REPLICATE that is used to specify Level 2 weights is not accurate. The documentation states that the variable must have positive integer values; however, this is not the case (Kathleen Kiernan, SAS Technical Support Statistician, personal communication, June 2012). The values of the variable may be positive Real numbers. Note when the REPLICATE variable contains non-integer values, the number of clusters reported in the output will be incorrect.

Bibliography

A. Agresti. *Categorical Data Analysis*. Wiley, Hoboken, NJ, 2 edition, 2002.

A. Agresti. *Introductory Categorical Data Analysis*. Wiley, Hoboken, NJ, 2 edition, 2007.

P. D. Allison. *Missing Data*. Sage, Newbury Park, CA, 2002.

S. R. Amer. Neural network imputation in complex survey design. *International Journal of Electrical and Electronics Engineering*, 3:52–57, 2009.

C. J. Anderson and L. Rutkowski. Multinomial logistic regression. In J. Osborne, editor, *Best practices in quantitative methods*, pages 390–409. Sage, Thousand Oaks, CA, 2008.

T. Asparouhov and B. Muthén. Mutlilevel modeling of complex survey data. In *Proceedings of the Joint Statistical Meeting*, pages 2718–2726. ASA section on Survey Research Methods, Seattle, WA, 2006.

Eugene Demidenko. *Mixed Models: Theory and Applications*. Wiley, Hoboken, NJ, 2004. ISBN 0-471-60161-6.

C. K. Enders. *Applied Missing Data Analysis*. Guilford Press, New York, NY, 2010.

L. Fahrmeir and G. Tutz. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, New York, NY, 2 edition, 2001.

L. Grilli and M. Pratesi. Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs. *Survey Methodology*, 30:93–103, 2004.

Jonathan Hartzel, Alan Agresti, and Brian Caffo. Multinomial logit random effects models. *Statistical Modelling*, 1:81–102, 2001.

Steven G. Heeringa, Brady T. West, and Patricia A. Berglund. *Applied Survey Data Analysis*. Chapman Hal/CRC, Boca Raton, FL, 2010. ISBN 978-4200-8066-7.

J.-S. Kim and C.M. Swoboda. Strategies for imputing missing values in hierarchical data: Multilevel multiple imputation. Paper presented at the *Annual Meeting of American Educational Research Association*, Vancouver, CA, 2012, April.

C. E. McCullagh and J.A. Nelder. *Generalized Linear Mixed Models*. Chapman and Hall, London, 2 edition, 1989.

D. McFadden. Conditional logit models analysis of qualitative choice behavior. In P. Aarembka, editor, *Frontiers of Econometrics*. Academic Press, New York, NY, 1974.

G. Molenberghs and G. Verbeke. *Models for Discrete Longitudinal Data*. Springer, New York, NY, 2005.

B. O. Muthén. Mplus technical appendices. Technical report, Muthén & Muthén, Los Angeles, CA, 1998–2004.

L. K. Muthén and B. Muthén. *Mplus User's Guide*. Muthén & Muthén, Los Angeles, CA, 6 edition, 1998-2010.

B. Peterson and F. Harrell. Partial proportional odds models for ordinal response variables. *Applied Statistics*, 39:205–217, 1990.

D. Pfeffermann, C.J. Skinner, D.J. Holmes, H. Goldstein, and J. Rasbash. Weighting for unequal selection probabilities in multilevel models (with discussion). *Journal of the Royal Statistical*

Society, Series B, 60:23–56, 1998.

D. D. Powers and Y. Xie. *Statistical Methods for Categorical Data Analysis*. Emerald Group, Bingley, UK, 1999.

S. Rabe-Hesketh and A. Skronkal. Mulitlevel modeling of complex survey data. *Journal of the Royal Statistical Society, Series A*, 169:805–827, 2006.

S. Rabe-Hesketh and A. Skronkal. *Multilevel and Longitudinal Modeling Using Stata, VolumneII, 3rd Edition*. Stata Press, College Station, TX, 2012. ISBN 978-1-59718-104-4.

S. W. Raudenbush and A. S. Bryk. *Hierarchical Linear Models*. Sage, Thousand Oaks, CA, 2nd edition, 2002.

J.P. Reiter, T.E. Raghunathan, and S.K. Kinney. The importance of modeling smapling design in multiple imputation for missing data. *Survey Methodology*, 32:143–149, 2006.

D. Rubin. *Multiple Imputation for Nonresponse in Surveys*. Wiley, NY, 1987.

L. Rutkowski, E. Gonzalez, M. Joncas, and M. von Davier. International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher*, 39:142–151, 2010. doi: 10.3102/0013189X10363170.

SAS Institue Inc. *The SAS System*. SAS Institue, Cary, NC, version 9.3 edition, 2011a.

SAS Institue Inc. *SAS/STAT 9.3 User's Guide: The GLIMMIX Procedure*. SAS Institue, Cary, NC, version 9.3 edition, 2011b. ISBN 978-1060764-935-9. URL `http://support.sas.com/documentation/onlinedoc/stat/930/glimmix.pdf`.

J. L. Schafer. *Analysis of Incomplete Missing Data*. Chapman & Hall, London, 1997.

G.S. Self and K.Y. Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82: 605–610, 1987.

Y. Shin and S. W. Raudenbush. Just-identified versus overidentified two-level hierarchical linear models with missing data. *Biometrics*, 63:1262–1268, 2007.

Y. Shin and S. W. Raudenbush. A latent cluster-mean approach to the contextual effects model with missing data. *Journal of Educational and Behavioral Statistics*, 35:26–53, 2010.

A. Skrondal and S. Rabe-Hesketh. *Generalized Latent Variable Modeling*. Chapman Hal/CRC, Boca Raton, FL, 2000.

T. A. B. Snijders and R. J. Bosker. *Multilevel Analysis*. Sage, Thousand Oaks, CA, 2 edition, 2012.

D.O. Stram and J.W. Lee. Variance components testing in the longitudinal mixed effects model. *Biometrics*, 50:1171–1177, 1994.

C. M. Swoboda. *A new method for multilevel multiple imputation*. Unpublished doctoral dissertation, University of Wisconsin–Madison, 2011.

S. Van Buuren. Multiple imputation of multilevel data. In J. Hox and K. Roberts, editors, *Handbook of Advanced Multilevel Analysis*. Taylor & Francis, New York, NY, 2011.

S. Van Buuren. *Flexible Imputation of Missing Data*. Chapman Hall/CRC, Boca Raton, FL, 2012.

Geert Verbeke and Geert Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer, NY, 2000. ISBN 0-387-95027.