Chapter 5

Reviewers’ Comments

The four outside reviewers’ comments covered three broad areas: the general approach taken for this analysis, the specification of the variables used, and the interpretation of the results. Their comments are summarized here.

Bonnie S. Billingsley, Ed.D., Associate Professor
Virginia Tech, College of Education

General Approach

It is obvious from the review of theories in Chapter 3 that the authors have a good grasp of the teacher retention literature. They review several different theories and a broad range of variables that influence retention, and rely upon the writings of Grissmer and colleagues extensively, which makes sense given their recent contributions to the knowledge base in this area. The hypotheses posited for the multivariate analysis are consistent with the theories proposed in Chapter 3 and appear to be logical. Many of the variables have been included in previous research studies on attrition/retention and commitment. Some of the hypotheses have strong support based on the theories reviewed and previous research (e.g., age, gender).

The first and second models include different samples, recognizing that attrition factors are likely to differ for subgroups of teachers. I would also suggest investigating an “early career” model and including a sample of stayers and leavers with less than 10 years of experience (e.g., under a certain age, such as 35), since this is a likely-to-leave group. The proportion of teachers leaving will also be higher among this group.

Specification of Variables

The authors have included the major groups of variables needed to investigate attrition among teachers. There are other variables I would suggest including, but some of these are not in the SASS database. Others are suggested below.

The relatively large number of variables and their groupings complicate and confound the analysis. The authors do not provide a rationale for the groupings of variables, and it is not clear why some of the variables are grouped the way they are. For example, the demographic variables include variables that are not demographic measures (e.g., subject level taught). Also, the fact that the different groups are entered using a manual stepwise approach implies that the specific composition of the groups influence the various forms (specifications) of the model and consequently influence the statistical results.

Perhaps consider reducing the number of variables using either principal components or factor analysis to identify more defensible variable groups. These can be used in the software analysis as conceptual dimensions that reflect the meaningful information contained in the entire set of variables. In effect, the conceptual dimension “indexes” would reduce the number of variables required in the analysis. In their current form, the models are likely to include too many variables grouped in a questionable manner that confound the analysis through the application of the manual stepwise approach.
I would use a conceptual framework for teachers' career decisions based on a review of attrition/retention literature to evaluate the specific independent variables included in the models.58

Professional qualifications. Several variables that relate to teachers' professional qualifications are included in the analysis such as: highest degree earned, whether or not they are teaching in the field for which they are best qualified, whether or not they are certified in their primary assignment field, and teaching experience. Although a variety of other professional qualifications might be considered, they are not available on the SASS database. However, it would be worth considering whether there is a good proxy for "entry path" (teachers entering through traditional versus nontraditional routes).

Work conditions and work rewards. Prior research suggests that work conditions and rewards are associated with teacher retention. The authors discuss several "work condition/work reward" hypotheses. Some of the variables included in this analysis are more "behavioral" (e.g., number of students/class, student/teacher ratio, high minority enrollment, urbanicity, school size, teaching income, merit pay), while others might be considered "affective reactions" to the conditions of teaching (e.g., satisfaction with salary, satisfaction with teaching). It seems logical that the behavioral variables might be predictive of the affective measures (e.g., high class sizes predictive of lower job satisfaction).

Some of the student variables are likely to be highly correlated (high minority enrollment and free lunch eligible as well as number of students/class and student/teacher ratio). Perhaps high influence over school policy and high control over classroom practices could be combined into one "control" variable.

Another problem in evaluating these variables is that the reader is not provided with a description of the measures, and no reliability coefficients are given. Section 5 of the questionnaire contains many work-related items. However, I could only guess which items made up job satisfaction. Also, I could not determine whether the variables listed under item 31 (school problems) were included in the analyses. I would use factor or principal components analysis as outlined above to identify defensible work-related variables from those included in the SASS database. I would also include (a) support variable(s) (e.g., administrative, colleague, and parents) in the analyses, given the fact that previous research suggests an association between support and attrition. Other work-related variables of interest include opportunities for professional development, workload, and school climate. I would also consider including grade level in the analyses since a number of researchers have found that grade level taught has been related to attrition, with secondary teachers leaving sooner than elementary teachers.59

Personal factors. Appropriate teacher characteristics (e.g., age, race-ethnicity, gender, marital status, children, high family income) have been included in the models.

**Interpretation of Results**

**Multicollinearity.** It was asserted that almost all of the correlations among the independent variables were "very low." Yet WESLOG could not compute an F value for the public school sample due to a singular matrix. Such singularity generally implies the presence of at least two redundant variables. I suggest that a complete correlation matrix be provided. Further, it would be helpful to evaluate a more comprehensive set of multicollinearity diagnostics such as variance inflation factors, the eigenvalues and eigenvectors of the correlation matrix, or both. The singular matrix generated by WESLOG is a signal that such additional analysis of the specification is appropriate.

**Goodness-of-fit.** The R2s do not inspire much confidence as reported. I suggest re-evaluating the goodness-of-fit of the models in light of the fact that dichotomous dependent variable models are unlikely to produce R2s close to 1. Research by Morrison explains why it is important in interpreting logit models to recognize that the upper bound for R2 is probably significantly less than 1.60 I suggest using Morrison's approach to estimate the upper bound for R2 under the assumption that the predicted probabilities follow a beta distribution. An "effective" R2 can be estimated as the ratio of the observed-to-the-upper bound for R2. The observed (empirical) R2 is computed as: (model $\chi^2 - 2k$)/(-2L(0), where k is the number of variables omitting the intercept, and L(0) is the maximum log-likelihood including only the intercept. This analysis will add a more meaningful indication of goodness-of-fit to the analysis. Further, I suggest reporting and evaluating the number of correctly classified observations.

**Interpretation of the estimated coefficients.** I suggest reporting and interpreting the logit regression elasticity coefficients rather than the simple coefficients. Define Prob ($Y_i = 1$) as the probability that teacher i will leave teaching; $Z_i$ as the arbitrary index used to ensure that the predicted probabilities reside in the unit interval for all Xs (the vector of explanatory variables); and $f(Z)$ as the value of the logistic density function for each possible value of the $Z_i$ index. Thus, the elasticity coefficients can be an estimate for each explanatory variable as follows:

$$\frac{\delta \text{Prob} (Y_i = 1) / \delta X_j}{\text{Prob} (Y_i = 1)} = \frac{X_j}{f(Z_i)} \frac{\beta X_j}{\text{Prob} (Y_i = 1)}$$

I suggest using the mean value of each explanatory variable and the mean predicted probability of membership in class 1.61 The elasticity coefficients are easier to interpret than the current statistics.

**Summary**

Overall, the lack of meaningful results are possibly due to several factors, including: 1) the extreme split between leavers and stayers as you suggest; 2) the use of variables that at least appear to be highly correlated; 3) the use of too many dummy variables (it would be preferable to use continuous variables); and 4) the possible lack of good work-related measures (e.g., school climate, administrative support, etc.), although it was difficult to evaluate these variables due to insufficient description.


General Approach

In summary, I believe that the technical procedures to handle the complex nature of the data are state-of-the-art. My concern with the manuscript lies in the implementation of the logistic regression modeling procedure itself.

The discussion of the use of SAS versus WESLOG is sound. Landis et al. discuss this approach of using standard linear modeling software (with weights) in comparison to a procedure that accounts for the complex sample design, and advocate the approach the authors have used as a sound practical procedure.62

Restricting the analysis to voluntary leavers seems sound, with one exception. One aspect of teaching that is attractive to many is the fact that it is a career in which it is relatively easy to regain employment after a move (due to a spouse's relocation, for example). Thus, in a sense, such a move is voluntary—if the teacher were only able to obtain work in a few specific locations, this might well affect the decision of the family to relocate. It might be useful to develop models with and without movers, both for this reason and because at a higher level (e.g., national) movers do not constitute attrition at all.

The decision to model public and private teachers separately, rather than just to include sector as an independent variable, may have been a mistake. Differences between public and private attrition rates might be explained by differences in the age and sex distributions of these two groups of teachers, for example, and it would not be possible to discern and establish this confounding relationship with separate models for public and private school teachers. One could learn a lot by discovering which other factors have significant interaction with public/private and also by being able to measure and test the effect of public/private school, controlled for other factors. If there were many interactions, or if different data were available for public versus private teachers (e.g., district level data), then the use of distinct models would be appropriate; however, these issues need to be investigated more fully. This also raises the question about movers between public and private school systems. To what extent do teachers leave public schools to teach in private schools and vice versa?

Using multi-level data is not as much of a problem as it appears to be in this report, because the one-level model analysis used takes into account the hierarchical nature of the data. It is true that the models available are more restrictive. One cannot measure interactions between random effects at one level and fixed effects at another level, for example; they must be assumed to be 0. Also, one cannot partition the variance of the random components into, for example, school and teacher components. However, the inference about the parameters in the model will be correct and will not suffer from the problems of incorrect inference that occur if one simply attaches school-level variables to teachers and then analyzes teacher data ignoring the clustering of teachers in schools. The analyses would be equally valid if there were many teachers per school and district, but in that case, an alternative methodology (HLM) would be available.

To illustrate these points, consider a simple HLM:

\[
\text{logit } (P) = \beta_0 + \beta_1 X + \varepsilon \\
\beta_0 = \gamma_{00} + \gamma_{10} Y + \delta \\
\beta_1 = \gamma_{01} + \gamma_{11} Y + \theta;
\]

where \( P \) is the probability of staying; \( X \) is some teacher characteristic; \( Y \) is a school characteristic; and \( \varepsilon, \delta, \text{ and } \theta \) are random terms. This model can be expressed as:

\[
\text{logit } (P) = (\gamma_{00} + \gamma_{10} Y + \delta) + (\gamma_{01} + \gamma_{11} Y + \theta) X + \varepsilon \\
= \gamma_{00} + \gamma_{01} X + \gamma_{10} Y + \gamma_{11} X Y + \theta X + (\delta + \varepsilon).
\]

Provided that \( \sigma_0^2 = 0 \) and that one is not concerned about establishing the relative magnitude of \( \sigma_Y^2 \) and \( \sigma_\varepsilon^2 \), then this model is reduced to a single (teacher) level model, with the school characteristic \( Y \) treated as a teacher-level variable. WESLOG can give correct inference and estimates for the parameters \( \gamma_{00}, \gamma_{01}, \gamma_{10}, \text{ and } \gamma_{11}. \) The weakness of this approach is the reliance for its validity on the assumption that \( \sigma_0^2 = 0. \)

**Specification of Variables**

The factors “high satisfaction with salary” and especially “high satisfaction with teaching” are co-outcomes or at least intervening variables. One cannot change a person’s satisfaction with teaching without changing some other factors such as salary or pupil/teacher ratio. Thus, before including such factors in the model, I think that it is necessary to ensure that they capture an important component not reflected in the other variables, and that at least there are some ideas or theories as to what the (unmeasured) underlying factors are. It is not very useful to learn that the main reason teachers left teaching was because they did not like it.

The idea of considering continuous variables as both continuous and dichotomous is sound. It enhances the ability to detect continuous variables that have threshold or ceiling effects.

**Interpretation of Results**

\( R^2 \) meaning. My experience leads me to endorse the discussion about the inappropriateness of the \( R^2 \) statistic for logistic regression, and I think that the discussion about this is sound. I also agree that the estimated probability of a group’s leaving can be computed using the formula \( P = e^L/(1+e^L) \). With the variance-covariance matrix, one can derive large-sample confidence intervals for \( L \) in each case. Using the fact that \( P \) is a monotone function of \( L \), one can derive an asymmetric large-sample confidence interval for \( P \). That is, if the confidence limits for \( L \) are \( L_1 \)
and $L_u$, (lower and upper, respectively), then confidence limits for $P$ are $e^{L_L} / (1 + e^{L_L})$ and $e^{L_U} / (1 + e^{L_U})$, respectively.

**Statistical significance of the model.** If the model lacks significance overall using SAS, then might it not be concluded that none of these variables is worth pursuing further, even though some are individually significant using WESLOG (and SAS)? Otherwise, why worry about the overall test of fit at all?

I am very puzzled that with so many significant terms in the model, the overall model fit was not significant. Are the results being interpreted correctly? For instance, in the Private School Groups I, IV example, the overall model was highly significant with $p < 0.001$ ($F=6.55$ with 10, 39 degrees of freedom). Alternatively, perhaps this is a result of including so many terms in the model. Did you consider using a stepwise procedure because of the large number of variables involved? I am also somewhat surprised that the results overall are not significant for public schools, yet are for private schools. Presumably the sample sizes are smaller for private schools, and the sizes of the effects of the significant variables do not look very different. I wonder if the weights within each group (public and private) have been scaled so that they have a mean of 1.0 in each case. However, if they have not, I would expect the result to understate significance for private school teachers and overstate it for public school teachers, since my guess is that SASS oversamples private school teachers. (This is indicated by the sample sizes.)

**Singular matrices.** Korn and Graubard discuss the issue of singular matrices in analyzing complex survey data.63 I do not think this is a serious problem here as the approach of using SAS to assess overall model fit, checking against WESLOG results for the comparability of individual parameter significance, is quite sound.

**Fay weights.** The WESLOG results can be corrected by simply multiplying all of the standard error estimates obtained from WESLOG by a factor of 2 (derived from the use of factors 0.5 and 1.5 in obtaining the Fay weights). This will eliminate some variables from tables 21 and 23 and probably change some **s to *. It will also make the results of these tables very similar to those in tables 22 and 24, again showing that you could reasonably proceed, on the basis of these analyses, with just using the SAS analyses.

**Summary**

The authors have done a good job of trying to deal with the technical issues involved in using logistic regression with complex survey data. However, I see three major problems. First, logistic regression just does not really explain why teachers leave teaching. The values in tables 22 and 24 are either close to 1 or else are extremely unreliable, as in the case of “male.” Second, no consideration has been given to possible interactions among the significant main effects. It seems highly plausible that such interactions might exist and might even substantially increase the explanatory power (fit) of the models. Finally, there is no discussion of the interpretation or plausibility of the final models. What is the overall message? Does it make sense, for example, in table 24 that “high satisfaction with salary” has a positive coefficient, indicating that those more satisfied are more likely to leave?

The discussion in Chapter 6 (Conclusions) of whether the sample of teachers who leave for career reasons is too small to be systematically different from other teachers gets to the heart of the problem in identifying the characteristics of a relatively rare group. It is unlikely that the

---

really important variables have been measured. The only real hope lies in looking at two-way and perhaps higher order interactions, which was not done.

Stanley Lemeshow, Ph.D., Professor and Chair, Biostatistics and Epidemiology
University of Massachusetts, Amherst

General Approach

My comments focus on the logistic regression modeling used in this report. I am not familiar with the WESLOG program, but appreciate the fact that it is trying to incorporate the proper statistical weights into the estimation process. My comments relate to building the logistic regression model with conventional software. I believe that if you were to find a satisfactory model using standard software first, and then refit the final model with WESLOG, the process could be considerably easier and more understandable than it currently is. Also, many of the basic procedures used in building logistic regression models are simply not set up for situations where the statistical weights are unequal for each subject. It would be important to assess the performance of the model with traditional software because it is not possible with the more specialized software. Procedures such as goodness-of-fit testing, analysis of logistic regression diagnostics, and area under ROC curves should not be omitted.

All the models are over-parameterized. There are simply too many nonsignificant variables included in each model. This is manifested by the matrix singularities and strange results in the odds ratios and associated confidence intervals. The fact that the models are full of nonsignificant independent variables has led to very unstable results. For example, table 18 demonstrates a very poor model. Notice the very large standard errors associated with the “less than BA/BS” category or the large standard error of the “Black, non-Hispanic” group relative to the beta in that group. These are indications that the model has fallen apart statistically. It is much better to build smaller models that behave in a stable manner, have statistically significant terms, and fit. My approach would have been to build smaller, significant, and well-fitting models with SAS, and then to refit those models with the WESLOG method incorporating jackknifing or BRR to corroborate the results and better estimate parameters.

I would concentrate more on assessing the performance of the models. With modern computer software it is easy to fit models to data. For example, anyone can fit a straight line to a set of X’s and Y’s, but the straight line is not always appropriate. This has to be checked with logistic regression, as well by computing goodness-of-fit tests and examining diagnostic statistics to see if there are any highly influential or poorly fit covariate patterns. More attention should be paid to these issues.

Specification of Variables

The use of dummy variables in logistic regression analysis is very important. The 0, 1 coding chosen for the dummy variables is fine if there is an appropriate reference group. An alternative would be to use +1, -1 coding to allow comparison of each region to the national average, rather than to a single reference group. This “deviation from means” type of coding is quite useful in some instances, as opposed to the more commonly used “reference cell” type of coding used in this report.

I would suggest that the scale of an independent variable X be checked before including it in the logistic model as a continuous variable. Inclusion assumes that the logit is linear in X, which may or may not be the case. This assumption must be tested. If it is not true, then the
variable can be categorized or transformed to make it more appropriate for inclusion in a logistic regression model.

The report mentions that bivariate tests were conducted relating each independent variable to the dependent variable. It is very important that this be done in order to select (from many potential predictor variables) those that are related to the outcome in a crude sense. If a variable is not related to the outcome in this simple screen, it probably will do more harm than good to include it in the model. The P values associated with these bivariate tests should be reported.

It is also mentioned that “the intercorrelations of the independent variables and the correlations with the dependent variable were examined for possible multicollinearity.” Multicollinearity in logistic regression manifests itself by large regression coefficients, large standard errors associated with these coefficients, and highly variable confidence intervals for odds ratios. Although very useful in multiple linear regression, computing the tolerance or other functions of intercorrelations between the independent variables is not useful with logistic regression.

Interpretation of Results

I object very strongly to the use of $R^2$ as a measure of performance of logistic regression models. It simply does not have the same interpretation as it does for linear regression and is not useful. Some people use the Pearson chi-square or the deviance chi-square to assess fit. This also is an incorrect procedure for the type of data being analyzed here.

I would strongly suggest that any models be evaluated with formal goodness-of-fit tests. These are easy to perform with standard statistical packages such as SAS, SYSTAT (LOGIT module), STATA, and others that have the Hosmer–Lemeshow procedure built in. Because WESLOG will not do this, I suggest building and assessing the models first assuming the sample is a simple random one, and then after you are satisfied that you have a good model, refitting the coefficients and standard errors in a manner that accounts for the complex nature of the sample.

While the Hosmer–Lemeshow test is not without imperfections, it is the best we have right now, and it will let you know immediately if your model is not reflecting the true outcome experience in the data. The likelihood ratio test is not a measure of fit. It simply tests the significance of terms in a model or between models.

There is some discussion about 2 x 2 classification tables. We believe that such tables are only of limited value if the objective of your model is to make a binary prediction for each subject. The model may have very poor fit and have good classification. First, you need to examine fit. Also, when you place a subject into a 2 x 2 classification table, you are losing a lot of information about that subject’s probability of having the outcome. That is, a subject whose probability is .02 is considered the same for purposes of classification as another subject whose probability is .48. Similarly, a subject whose probability is .48 is predicted to have an entirely different outcome than one whose probability is .52. This seems to me to represent a distortion of the meaning of probability. Confidence intervals could be calculated for each subject, but it is not clear what you would do with them.
Donald E. Ramirez, Ph.D., Professor  
Department of Mathematics, University of Virginia

General Approach

The study involves the binary response variable of teachers who are leavers or stayers. Such studies are modeled using the binomial distribution. The probabilities are functions of linear combinations of teacher and school characteristics. Since some of the characteristics are continuous, the general logistic regression model is appropriate.

The classical statistical models involve assumptions that do not reflect the complexity in the SASS and TFS surveys. Therefore, they require software that will take into account the complex sampling design. The purpose of the analysis was to assess the impact of possible explanatory variables on the binary response variable. Prentice and Pyke have shown the remarkable fact that in this type of study the estimators of the parameters for the explanatory variables in a logistic regression model are consistent, as is the sample information matrix. The Prentice–Pyke result does assume that the logistic model is correct. However, this has yet to be established in this study.

The number of variables should be reduced. Table 18 gives the overall view of the data with all 31 variables included. The improvement in the model is measured by the likelihood ratio test with $\chi^2_{31} = 35.33$. The next step should be a backwards elimination of nonsignificant variables. I would recommend deleting some of the nonsignificant variables, and recomputing the model and the $\chi^2$ goodness-of-fit test. I believe what will happen is that $\chi^2$ will be only slightly smaller than 35.33, but the degrees of freedom will change from 31 to 9 (4 for child, 4 for secondary, and 1 for free lunch). The test will then be very significant. SAS will do this automatically, but all dummy variables in a group should be retained if one is retained. The model can be further simplified by compressing the nonsignificant levels in a group into one level. To check that the deleted variables are really not significant, one can do a forward regression to compute the improvement with only that group of variables included.

Since the significant variables are either categorical or can be converted into such variables, I would also suggest categorical modeling. Since the dependent variable has only two responses, CATMOD and LOGIST will yield the same parameter estimates. The advantage of CATMOD is that the output will contain observed and predicted values for each cell. SPSSX uses the observed and expected frequencies to compute a Pearson goodness-of-fit test.

It would help to have a longer period of study to increase the number of leavers. If a 3-year interval were possible, the total sample would only increase slightly, but the number of leavers would be tripled.

---

Specification of Variables

The continuous variables appear to be monotone with respect to the binary response variable. Since retirees are not included in the study, the anticipated U-shaped relationship between leaving and age should have been eliminated.

Interpretation of Results

The $R^2$ is not appropriate for logistic models. As noted by Agresti, "Despite several attempts to define analogs of $R^2$ for models for categorical responses, no proposed measure seems as widely useful as the regression $R^2$."\textsuperscript{68} Agresti proposes a couple of measures that are essentially based on the maximized log likelihood. This study uses the (quite reasonable) approach of reporting the odds ratio.

The logit $L$ for a group is really $x'\beta$. When the variable is categorical, you can report $e^\beta$. However, when the variable is continuous, for example, $AGE*AGE$ in table 21, the value reported $e^\beta = 1.00$ is the odds for a teacher who is 1-year-old! For a teacher with AGE = 25, the odds is $\exp(.002*25^25) = 3.5 \neq 1.00$. The odds can be reported by centering the continuous variables at their mean value.

I do not believe that reporting the estimated probabilities $P = e^{x'\beta} / (1 + e^{x'\beta})$ will be very informative. However, it might be worth trying to do so since this is not difficult to calculate.\textsuperscript{69}

The stochastic assumptions for the classical logistic models are not satisfied by complex designs. In particular, the homogeneity of the population clusters will tend to increase the variance of the estimated parameters over the usual asymptotic estimators. It is interesting that this phenomenon did not occur, perhaps indicating that teachers are acting independent of the school effect.

\textsuperscript{68}Alan Agresti, \textit{Categorical Data Analysis} (New York: John Wiley, 1990), 110.
\textsuperscript{69}See SAS User's Guide Version 6, 1091.