DOCUMENT RESUME

ED 399 341 UD 031 306

AUTHOR Denson, Katy; Schumacker, Randall E.

TITLE Student Choices: Using a Competing Risks Model of Survival Analysis.

PUB DATE Apr 96

NOTE 46p.; Paper presented at the Annual Meeting of the American Educational Research Association (New York, NY, April 8-12, 1996).

PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS Academic Persistence; Coding; High Schools; *High School Students; *Predictor Variables; Probability; Profiles; *Regression (Statistics); Research Methodology; *Risk; Urban Schools; *Withdrawal (Education)

IDENTIFIERS *Dallas Independent School District TX; *Survival Analysis

ABSTRACT

By using a competing risks model, survival analysis methods can be extended to predict which of several mutually exclusive outcomes students will choose based on predictor variables, thereby ascertaining if the profile of risk differs across groups. The paper begins with a brief introduction to logistic regression and some of the basic concepts of survival analysis necessary to understand the competing risks survival analysis method. A data set is used to illustrate conducting the competing risks survival analysis, and results of that analysis for each competing risk and predictor variable are presented and interpreted. The procedure for conducting a competing risks survival analysis is similar to that of conducting a survival analysis with only one outcome. Data are prepared and coded in a like manner, and survival and hazard functions are interpreted by the same guidelines. However, in a competing risks survival analysis, the hazard probabilities for each competing outcome are recombined to create a complete profile of each risk for each time period in question. Data from the Dallas (Texas) public schools for 7,748 students (ninth graders in 1990-91) with no more than one of the defined competing risks (school leaving or completion) illustrate the analysis approach. Appendixes contain a Statistical Analysis System program for the analysis, withdrawal and dropout reasons for the student cohort, and a discussion of the transformation of data from person-data to person-period-data. (Contains 2 figures, 12 tables, and 33 references.) (SLD)

Student Choices: Using a Competing Risks Model of Survival Analysis

Katy Denson
Dallas Public Schools
Dallas, Texas

Randall E. Schumacker, Ph.D.
University of North Texas
Denton, Texas

Paper presented at the American Educational Research Association Conference
April, 1996

Student Choices: Using a Competing Risks Model of Survival Analysis

There is a nation-wide concern regarding the declining number of students who remain in school through graduation. For years, educational research has focused on the issue of student dropout (Pittman, 1995; McMillen, 1994; Fitzpatrick and Yoels, 1992; Ensminger & Slusarcick, 1992). It is well known that students today face challenges from various sources, and many are leaving high school before reaching graduation. Using survival analysis methods, researchers can investigate not only *if*, but *when*, students are most likely to leave school. Morrow (1986), in an attempt to standardize the analysis of school dropouts, recommended that the condition of "dropout" be defined more specifically. He identified several modes of exiting school that are typically classified as "dropout," including withdrawal to another district, expulsion, and dropouts who return to school at a later time.

Why Conduct a Competing Risks Survival Analysis?

By using a competing risks model, survival analysis methods can be extended to predict which of several mutually exclusive outcomes, or modes of leaving school, students will choose based on predictor variables, thereby ascertaining if the profile of risk differs across groups. "With competing risks survival analysis methods, any number of qualitatively different modes of exit can be modeled. By building hazard models of each of these events, we can better understand the different forces that drive different students to different ends" (Willett and Singer, 1991, p. 428).

Researchers have used numerous data collection and analysis methods to study, for example, students who drop out. Three data collection methods have been most prevalent, but there are specific problems associated with each of these methods. Retrospective studies gather information on a cohort of students to compute a dropout rate. Limitations of these studies include (a) pooling disparate groups of people, (b) excluding subgroups, (c) biased reporting of educational attainment levels, and most importantly, (d) they ignore the problem of censoring, and (e) they ignore the timing of *when* people dropped out (Willett and Singer, 1991).

Two-wave prospective studies are used by many school districts to calculate annual dropout rates. These dropout rates frequently use enrollment figures that aggregate students across many grade levels. Again, they provide little insight into who drops out and when. Willett and Singer (1991) provide a further description of the problems that arise from data aggregated across grade levels. Dropout rates by grade provide more information because they can identify when students are at the greatest risk of dropping out.

Although multiwave studies are more commonly used today, the calculated dropout rates can be misleading because each year's rate is based on a different cohort student group with differing social makeup and demographics. When only "end-product" statistics are calculated, such as total number of graduates, dropouts, or no-shows, those who are censored are left out of the model. The most significant limitation is the failure to identify *when* students are most likely to make a choice regarding their educational career. With a competing risks survival analysis model, researchers can simultaneously study all of the possible choices, reaching appropriate conclusions that answer questions about the *risk* of dropping out during specific time periods.

This paper begins with a brief introduction to logistic regression and some of the basic concepts of survival analysis necessary for an understanding of the competing risks survival analysis method. Following that is a description of the data set, how the competing risks survival analysis was conducted, and the results of the analyses for each competing risk and predictor variables. Finally, results of the six competing risks modeled by the procedure are interpreted.

Logistic Regression

The logistic regression model is formulated for use with interval level data on independent variables and dichotomous data on the dependent variable. The related logit model is more appropriate when both dependent and independent variables provide dichotomous or categorical data.

The Logistic Regression Model

A standard regression analysis of data with a single dependent variable, X, would yield a simple regression model $Y' = \alpha + \beta X$. This basic least-squares regression model is usually not suitable for dichotomously scored dependent variables because probabilities may fall below 0 and above 1. In logistic regression models, a curvilinear relationship, rather than a linear one, occurs because of the nature of the dependent variable coding. Therefore, a logarithmic transformation is necessary to linearize the logistic response functions, creating probabilities that fall in the range of 0 to 1 (Neter, Wasserman, and Kutner, 1989). Through this transformation, the logistic regression model can be expressed directly in terms of probability as:

$$p = e^{\alpha + \beta x} / [1 + e^{\alpha + \beta x}]$$

The logistic regression coefficient, β, can be interpreted as an effect on the odds. Taking the antilog of both sides of the logistic regression equation, the following is obtained:

$$\log[p/1\text{-}p] = e^{\alpha + \beta x} = e^{\alpha}(e^{\beta})^{x}$$

"The right-hand side of the equation has an exponential form that implies that every unit increase in X produces a multiplicative effect of e^{x} on the odds" (Agresti and Finlay, 1988, p. 485).

Logistic regression models are estimated using maximum likelihood rather than ordinary least-squares, as in linear regression. Wright (1995) states that "in logistic regression, the maximum likelihood criterion is generally used for selecting parameter estimates. The coefficients maximize that probability (likelihood) of obtaining the actual group memberships for cases in the sample. Thus, the logistic regression coefficients are known as maximum likelihood parameter estimates" (p. 225). This is done through an iterative process in which the computer program finds successively better approximations of the β values that satisfy the maximum likelihood equations.

Assumptions of the Logistic Regression Model

If specific assumptions about the population are met, maximum likelihood estimates of logit parameters should be unbiased, efficient, and normal with large enough data samples (Hamilton, 1992; Hildebrand, 1986). *First*, it is assumed that the random dichotomous variable takes the value 1 with probability P_1 and the value 0 with probability $P_0 = 1 - P_1$. *Second*, the outcomes must be statistically independent. In other words, a single case can be represented in the data set only once. *Third*, the model must be correctly specified so that it contains all relevant predictors and no irrelevant predictors. This specification assumption, however, is rarely met.

Fourth, the categories or outcomes must be *mutually exclusive* and *collectively exhaustive*. This means that a case cannot be in more than one outcome category at a time, and every case must be a member of one of the categories under analysis. *Fifth*, none of the X variables are linear functions of the others. Perfect multicollinearity makes estimation impossible, and strong multicollinearity makes estimates imprecise. *Finally*, because the standard errors for maximum likelihood coefficients are large sample estimates, the sample must be large. For most cases, a minimum of 50 cases per predictor variable is sufficient to test hypotheses involving the logistic regression coefficients (Hamilton, 1992; Wright, 1995).

Discrete-Time Survival Analysis

Researchers frequently wish to ask questions related to the timing of developmental or educational events that occur in various populations and the variables that impact these events. Events such as amount of time children spend in day care (Singer, Fosburg, Goodson, and Smith, 1978), teacher attrition (Murnane, Singer, and Willett, 1988, 1989), high school student dropout and graduation (Sween, 1989; Roderick, 1994), and doctoral program completion (Zwick and Braun, 1988) have been studied using survival analysis methods to answer, not just whether the event occurs, but when it is most likely to occur, and under what conditions. Survival analysis is unique in that it can handle both time-varying and time-invariant predictor variables and uses data from all observations, censored or uncensored. A case is considered to be censored if the event in question did not occur before the end of data collection.

The Survivor Function

The analysis begins with an examination of the survival probability function. This survivor function is a plot of the probability that an individual will remain in the risk pool as a function of time. The shape of the survivor function is very consistent — a negatively accelerating, monotonic extinction curve (Singer and Willett, 1991). At the beginning of a study, when all individuals are present, the survival probability is 1.00. As time passes, and individuals experience the event in question, the survival probability drops toward 0.0, though rarely reaching it because every case usually does not experience the event before data collection ends.

The Hazard Function

The hazard function has been called the "cornerstone" of survival analysis for three reasons: (1) it shows whether and, if so, when events occur, (2) information from both censored and uncensored cases is included, and (3) the sample hazard function can be computed for every time period under consideration, then plotted, to reveal variation in the timing of events (Singer and Willett, 1993). The hazard function mathematically registers changes in the slope of the survivor function, thereby allowing the researcher to identify high risk time periods. The higher the hazard, the higher the risk that the event will occur.

Statistical Models of Hazard

Relationships between entire hazard profiles and one or more predictors are hypothesized in the hazard models. The entire hazard function is the conceptual outcome, with other variables added as potential predictors of that outcome. "A population hazard model formalizes this conceptualization by ascribing the vertical displacement to the predictors in much the same way as an ordinary linear regression model ascribes differences in mean levels of any continuous noncensored outcome to predictors" (Willett and Singer, 1991, p. 416).

Because the variables included are measured at different levels, the hazard profiles must be transformed logarithmically to put all variables on the same level of measurement. (See Ferguson and Takane, 1989, for a discussion of acceptable transformations of data.)

Time is measured in discrete, rather than continuous, intervals so that a logistic transformation is appropriate. If *p* represents a probability, then logit(*p*) is the natural logarithm of *p/(1-p)*; so logit(*p*) can be interpreted as the conditional log-odds of the event in question occurring (Allison, 1984).

The Baseline Model. $\beta_0(t)$ is the baseline log hazard profile, and represents the values of the outcome without other predictor variables. The baseline equation can be expanded to account for specific measurements of discrete time intervals to

$$\text{logit}_e(h)_j = [\alpha_1 T_1 + \alpha_2 T_2 + \ldots \alpha_k T_k]$$

The alpha parameters are "multiple intercepts, one per time period" and represent the "baseline logit-hazard function because it captures the time-period by time-period conditional log-odds that individuals whose covariate values are all zero will experience the event in each time period, given that they have not already done so" (Singer and Willett, 1993, p. 167).

Adding Predictor Variables. As in multiple regression, the equation expands to include predictor variables that control for observed heterogeneity. The relationship of the log-transformed hazard profile to the predictor variable, X_1, is

$$\text{logit}_e(h)_j = [\alpha_1 T_1 + \alpha_2 T_2 + \ldots \alpha_k T_k] + \beta_1 X_1$$

Interaction terms can also be included in the hazards model. Cross-product terms are added to the main effect models in the same manner in which interactions are examined in multiple regression. The β parameters measure the amount of "vertical shift" in log-hazard per unit difference in the predictor variables.

Assumptions for the Discrete-Time Hazard Model

Having postulated the discrete-time hazard model using logistic regression, Singer and Willett (1993) point out three assumptions. The assumptions are (1) linearity, (2) no unobserved heterogeneity, and (3) proportionality. Linearity is similar to linearity in regression, with the addition that "vertical displacements in logit hazard are linear per unit of difference in each predictor" (Singer and Willett, 1993, p. 182).

No unobserved heterogeneity refers to the assumption that the inclusion of predictors in the model accounts for all of the error. Thus, it becomes very important to choose the correct predictors and not omit relevant predictors.

As described in Cox's model (Cox, 1972), proportionality refers to the assumption that logit hazard profiles of various predictors maintain the approximate shape of the baseline profile, but shift it up or down, depending upon the sign of the β value. If data are not checked for nonproportionality, results may be biased. Other event history analysis models make no allowance for the violation of this proportionality assumption, although nonproportionality does occur frequently.

In discrete-time survival analysis, it is relatively easy to ascertain whether the proportionality assumption has been violated. Singer and Willett (1991) have developed a SAS program creates new dummy variables that reflect the effects of the predictors over time. (See Appendix A.) These new variables are cross-products between the time indicators (α_1T_1, α_2T_2, etc.) and the predictors. This procedure allows the data to be checked both graphically and statistically. A visual examination of graphs of the hazard functions for Y=1 and Y=0 will indicate whether there is a near-proportional distance between the two lines. Significant differences between the profiles can be checked statistically by consulting a Bonferroni table to evaluate critical F values (Denson and Schumacker, 1994).

Method

The procedure for conducting a competing risks survival analysis is similar to that of conducting a survival analysis with only one outcome. Data are prepared and coded in a like manner and survival and hazard functions are interpreted by the same guidelines. However, in a competing risks survival analysis, the hazard probabilities for each competing outcome are recombined to create a complete profile of each risk for each time period in question. The use of a competing risks model that focuses on the *combination* of events, rather than the analysis of each event separately, gives a more realistic picture of the pattern of choices.

The Data Set

Data were obtained from the database of the Dallas Public Schools on a cohort of students who began the ninth grade for the first time in the 1990-91 school year. These students were followed over the next four school years. Six competing risks were identified and coded on the database: (a) withdrawal from school for reasons identified as legal by the State, (b) dropping out of school, (c) graduation, (d) still enrolled in school after four years, (e) no-show status, and (f) unknown outcome. (For a list of specific reasons for leaving school and their coding as either withdrawal or drop out, see Appendix B.) Cases were eliminated from the data set if any of the following occurred:

1) Multiple drops or withdrawals during the four years,

2) Incomplete data from the database, such as no withdrawal date or reason, or

3) Withdrawal coding did not match a *known* outcome.

After removing the above noted cases, a total of 7, 748 students with no more than one of the competing risks remained. As can be seen in Table 1, almost half (47.8%) of the students had graduated by the end of the 1994 school year, 20% had dropped out, 15% had withdrawn, 7% either had been identified as having no known outcome or were still enrolled at the beginning of the 1994-95 school year, and 2% had been identified as no-shows.

Students were also coded respective to their status on the following variables, previously identified in the literature as predictors of dropout status: (a) gender (Lakebrink, 1989), (b) ethnicity (Rumberger, 1995; Miller, 1989), (c) special education enrollment (Kortering and Blackerby, 1992), (d) identification as limited English proficient (LEP) (Watt and Roessingh, 1994), (e) retention at some time during grades 1-8 (Nason, 1991; Roderick, 1994), and (f) overage relative to their class members (Orr, 1987). Numbers and percentages of students in each of the predictor categories are also included in Table 1.

Table 1

Demographic Information for Data Set

Censors/Predictors	N	%
Outcomes		
Withdrawal	1,139	15.3
Dropout	1,512	20.3
Graduation	3,556	47.8
Still Enrolled	568	7.6
No-Show	127	1.7
No Known Outcome	530	7.1
Gender		
Male	3,682	49.5
Female	3,751	50.5
Ethnicity		
Anglo	1,312	17.6
African American	3,596	48.4
Hispanic	2,355	31.7
Asian	169	2.3
Other Predictors		
Limited English Proficient	637	8.6
In Special Education	506	6.8
Retained in Grades 1 - 8	1,110	14.9
Overage	2,775	37.3
Total Population	7,432	100.0

Preparing the Data Set

Preparing the data set for discrete-time survival analysis using logistic regression involved coding the predictor variables dichotomously as [0,1], "0" indicating the absence of and "1" indicating the presence of the variable value. Because the entire data set was used for the separate analysis of each outcome, dummy variables were created indicating which of the six outcomes the student was coded. This modification of the definition of censoring allowed for the analysis of the competing risks. In this particular analysis, there were no time-varying variables, although discrete-time survival analysis handles the inclusion of both time-varying and time-invariant variables quite easily.

Before using logistic regression to conduct a discrete-time survival analysis, the data structure was transformed from the standard one-person, one-record data set (the person-data set) into a one-person, multiple period data set (the person-period data set) (Singer and Willett, 1991). Singer and Willett's (1991) SAS program was used to array the data in such a fashion. (See Appendix C for an example of the transformation.) For this analysis, there were eight time periods, corresponding to the naturally occurring eight semesters in the four school years (1990-91, 1991-92, 1992-93, and 1993-94). The records in the reconstructed person-period data set indicated what happened to each student during each discrete-time period when the outcomes of interest could have occurred, until one did occur, or until data collection ended (whichever came first).

The reconstructed data set yielded one record per semester per person. Each person-period record contained period-specific values of 19 different types of predictors, as well as several other variables used for identification (ID), specification of the last period the student was enrolled (LASTPD8), and the student's mode of exiting school (CENSOR). Table 2 contains the name, the dummy variable name that was created (if necessary), and the meaning of each variable. An annotated version of Willett and Singer's SAS program, modified to conduct a discrete-time competing risks survival analysis, can be found in Appendix A. This program also fits the model and reconstructs fitted hazard and survival plots from parameters estimates.

Table 2

Variables Included in the Discrete-Time Competing Risks Survival Analysis

Variable Name	Dummy Variable Name	Meaning of Variable
		Input Variables
ID	-	Student identification number assigned by District
SEX	-	Gender
ETHNIC	-	Ethnicity
LEP	-	Limited English Proficiency status
SPED	-	Special Education status
RETAIN	-	Retention status
OVERAGE	-	Overage status
LASTPD8	-	Refers to the last semester the student was enrolled
CENSOR	-	Indicates student's mode of exiting school
		Dummy Censor Variables
WD	-	Indicates student withdrew
DROP	-	Indicates student dropped out
GRAD	-	Indicates student graduated
STILLIN	-	Indicates student was still enrolled
NOSHOW	-	Indicates student was a no-show
NOKNOW	-	Indicates database had no known outcome
		Dummy Ethnicity Variables
ANGLO	-	Indicates student is Anglo
BLACK	-	Indicates student is African American
HISP	-	Indicates student is Hispanic
ASIAN	-	Indicates student is Asian
		Dummy Variables
OCCASION	E1 - E8	Specifies discrete-time interval to which record refers
SEXTIME	SX1 - SX8	Reflects the effect of gender over time
ETHTIME	ETH1 - ETH8	Reflects the effect of ethnicity over time
LEPTIME	L1 - L8	Reflects the effect of LEP status over time
SPETIME	SP1 - SP8	Reflects the effect of special education over time
RETTIME	R1 - R8	Reflects the effect of retention over time
OVRTIME	O1 - O8	Reflects the effect of being overage over time
ANGTIME	AN1 - AN8	Reflects the effect of being Anglo over time
BTIME	B1 - B8	Reflects the effect of being African American over time
HISTIME	H1 - H8	Reflects the effect of being Hispanic over time
ASTIME	AS1 - AS8	Reflects the effect of being Asian over time

Procedure for Conducting a Competing Risks Survival Analysis

Six separate survival analyses were conducted using the entire data set, one for each outcome that was analyzed. Through dummy coding, students who did not experience the outcome in question were treated as censored. (See dummy censor variables in Table 2.) A total of 114 hazard profiles were created by calculating hazard models for the baseline (1 analysis), each of the predictor variables (9 analyses), and the cross-products of each predictor with time (9 analyses) for each outcome.

After identifying the predictors of hazard for each outcome separately, the risk profiles for each outcome were recombined to create an overall risk for all events taken together. Hazards for each competing outcome were also combined for each predictor variable to compile a complete risk profile for each of the predictor variables and for the effect of each predictor variable across time. Although not discussed in this paper, the last set of hazard profiles could be used to check the proportionality assumption.

Results and Discussion

Baseline Models

The baseline models represent the values of the outcome without other predictor variables. Maximum likelihood estimators were not calculated for every time period for the outcomes of graduation and still enrolled because those events models could not occur in every time period. To make time periods more meaningful, they will henceforth be indicated by the grade and semester they represent. In other words, time period 1 will be labeled 9(1), indicating first semester of the 9th grade; time period 2 will be labeled 9(2), meaning second semester of the 9th grade, and so forth.

Attempting to determine estimators for time periods where no event occurs causes a quasicomplete separation in the data points. Menard (1995) cites several causes of quasicomplete separation: (a) collinearity in the independent variables, (b) zero cell count, which occurs frequently when using categorical variables, and (c) near perfect or perfect prediction of the dependent variable with a set of predictors. As graduation could only occur during the last three periods (second semester of the 11th grade and both semesters of the 12th grade) and still being enrolled could only occur after the last time period (second semester of the 12th grade), attempting to compute maximum likelihood estimates with zero cell counts causes quasicomplete separations. Likewise, there were no unknown outcomes or no-shows coded for time periods one and two (both semesters of the 9th grade). These time periods were, therefore, excluded from the analyses to eliminate the zero cell count and allow the determination of maximum likelihood estimates for those outcomes (Menard, 1995). The hazards for each baseline model are listed in Table 3.

Table 3

Baseline Hazards for Each Competing Risk Model

	Time Periods							
Model	9(1)	9(2)	10(1)	10(2)	11(1)	11(2)	12(1)	12(2)
Withdrawal	.0062	.0072	.0475	.0200	.0414	.0176	.0321	.0277
Dropout	.0424	.0658	.0129	.0255	.0148	.0242	.0392	.0194
Graduate	-	-	-	-	-	.0069	.0048	.8069
Still Enrolled	-	-	-	-	-	-	-	.1309
No-Show	-	-	.0005	.0015	.0026	.0041	.0147	.0012
Unknown Outcome	-	-	.0084	.0131	.0120	.0150	.0376	.0141

Note. "-" Indicates that no maximum likelihood estimates were calculated for these analyses due to zero cell counts.

Hazards can be directly interpreted as probabilities that the event will occur in that time period. For example, there is a 0.7% chance that any student who is still in the risk pool by the second semester of the 11th grade will graduate, a 0.5% chance that any student who is still in the risk pool will graduate after the first semester of the 12th grade, and an 81% chance that any student who remains in school through the end of second semester of the 12th grade will graduate. An examination of the hazards for each competing risk across time periods reveals that students are always at the greatest risk of either withdrawing or dropping out until the end of the senior year, when graduation is most likely.

Main Effect Models

For each competing outcome, the addition of predictor variables usually enhances the ability to predict the outcome. This is assessed through the use of the likelihood ratio chi-square test, a procedure very similar to testing for the significance of increments of R^2 when additional variables are added to a multiple regression equation. In logistic regression, the log-likelihood is the criterion for selecting parameters, and when multiplied by -2 has approximately a chi-square distribution. Larger values indicate a *worse* prediction of the dependent variable (Menard, 1995). To compare the fit of the two models, the -2LL (twice the positive difference between their log-likelihoods) is calculated and compared. In most cases the associated degrees of freedom will be the difference between the number of variables in the two models.

This procedure was employed to compare the main effect model fit statistics for each variable for each competing risk. The -2LL for the base model was subtracted from the -2LL of each predictor model. The -2LL for each model and the change in the -2LL reflecting the addition of each predictor can be found in Table 4.

From an examination of Table 4, it can be seen that some variables caused a greater change in the -2LL chi-square statistic than others. For all six competing outcomes, the inclusion of information regarding retention and overage status produced the most change in the -2LL. Other predictor variables, such as gender, were more informative for some outcomes than others. The effect of gender contributed a larger change for the outcomes of graduation or still enrolled, but little for the outcomes of withdrawal or dropout.

Table 4

Comparison of Main Effect Model Fit Statistics for Each Variable for Each Competing Risk

	Model									
Statistic	Base	Gender	Anglo	Afro Am	Hispanic	Asian	LEP	SpEd	Retain	Overage
					Withdrawal					
-2LL	10311	10297	10160	10242	10308	10311	10310	10311	10241	10142
Change in -2LL		14	151	69	3	0	1	0	70	169
					Dropout					
-2LL	12989	12969	12987	12946	12955	12989	12971	12977	12825	12317
Change in -2LL		20	2	43	34	0	18	12	164	672
					Graduate					
-2LL	50702	8595	45002	30415	38602	49952	47380	48108	44632	37202
Change in -2LL		42107	5700	20287	12100	750	3322	2594	6070	13500
					Still Enrolled					
-2LL	63370	4824	54066	34987	46838	62102	59109	60064	56482	46279
Change in -2LL		58546	9304	28383	16532	1268	4261	3306	6888	17091
					No-Show					
-2LL	21669	2075	18295	12541	15581	21265	20088	20362	18977	14835
Change in -2LL		19594	3374	9128	6088	404	1581	1307	2692	6834
					Not Known					
-2LL	25451	6502	22411	16810	20614	25044	24304	23410	24411	20022
Change in -2LL		18949	3040	8641	4837	407	1417	2041	1040	5429

Models Including the Interaction with Time

To maintain the assumption of proportionality, logit hazard profiles of various predictors must retain the approximate shape of the baseline profile. Frequently, predictors' hazards do not simply shift the baseline up or down, but actually change the shape. If the effect of a predictor varies over time, there is an interaction between that variable and time, and a nonproportional hazard model should be used. Because main effect models constrain the hazard profiles to be proportional (Singer and Willett, 1993), the inclusion of the cross-products of the predictors with time in the regression equation *may* reveal a truer reality. Singer and Willett (1993) warn that "serious consequences await those who blindly fit proportional-odds models without examining the tenability of the assumptions" (p.189).

In this study, some of the predictor variables appear to interact with time, thus violating the proportionality assumption. When appropriate, models that include an interaction with time, rather than main effect models, are interpreted. As with main effect models, the likelihood ratio chi-square test can be used to assess the fit of the models. The -2LL for the base model was subtracted from the -2LL of each predictor model for each competing outcome. This procedure produced the information found in Table 5, which lists the -2LL and the change in the -2LL for each predictor crossed with time.

Table 5

Comparison of Interaction Model Fit Statistics for Each Variable for Each Competing Risk

	Model									
Statistic	Base	Gender	Anglo	Afro Am	Hispanic	Asian	LEP	SpEd	Retain	Overage
				Withdrawal						
-2LL	10312	10286	10138	10229	10297	-	10301	10306	10224	10100
Change in -2LL		26	174	83	15		11	6	88	212
				Dropout						
-2LL	12989	12960	12928	12873	12942	29292	12961	12970	12726	12246
Change in -2LL		29	61	116	47	16303	28	19	263	743
				Graduate						
-2LL	50702	50607	50665	50697	50676	-	-	50663	50508	50575
Change in -2LL		95	37	5	26			39	194	127
				Still Enrolled						
-2LL	63370	63305	63327	63369	63335	63367	63332	63326	63265	63309
Change in -2LL		65	43	1	35	3	38	44	105	61
				No-Show						
-2LL	21668	21666	-	21659	21662	-	-	-	21637	21649
Change in -2LL		2		9	6				31	19
				Not Known						
-2LL	25451	25419	25430	25372	25373	-	25392	25436	25326	25252
Change in -2LL		32	21	78	78		59	15	125	199

Note. A "-" Indicates that no maximum likelihood estimates were calculated for these analyses due to a quasi-complete separation in the data points.

When comparing the information from Table 4 with that of Table 5, it can be seen that for the outcomes of withdrawal and dropout, the inclusion of the cross-products of the predictor variable and time in the regression equation caused a greater change in the -2LL than the inclusion of the predictor variable alone. This reinforces the need to interpret the interaction models, rather than the main effect models. For the other four outcomes, the main effect models produced the greater change in the -2LL. In order for maximum likelihood estimators to be calculated, in other words, to avoid a quasicomplete separation in the data points due to zero cell counts, the SAS program (Appendix A) was altered to include only those time periods in which the event could have occurred. Even with this modification, the inclusion of the predictor variables LEP and special education status in the logistic regression equation caused a quasicomplete separation in the data points, and no maximum likelihood estimates were calculated. Logically, if the event could not have occurred in any one of the eight time periods, the models that included an interaction with time would not be appropriate to interpret.

Interpretation of the Six Competing Risks

Withdrawal

As indicated from the changes in -2LL (Table 5), the model including the interaction with time is the appropriate one to interpret. The hazard probabilities found in Table 6 are a result of this interaction. Asian students withdrew only between time periods 10(1) and 12(1). Consequently, there are no estimators for periods 9(1) and 9(2).

Table 6

Hazard Probabilities of Withdrawing in Each Time Period
by Predictor Variables

	Time Periods							
Variable	9(1)	9(2)	10(1)	10(2)	11(1)	11(2)	12(1)	12(2)
Gender								
Male	0.05	0.02	0.07	0.03	0.04	0.04	0.05	0.05
Female	0.01	0.01	0.05	0.02	0.04	0.02	0.04	0.03
Ethnicity								
Anglo	0.01	0.02	0.10	0.03	0.09	0.03	0.04	0.03
Afro Am	0.00	0.01	0.03	0.02	0.03	0.01	0.03	0.03
Hispanic	0.00	0.00	0.04	0.02	0.04	0.02	0.03	0.02
Asian[a]	-	-	0.06	0.01	0.05	0.02	0.05	0.04
LEP Status								
EP	0.01	0.01	0.05	0.02	0.04	0.02	0.03	0.03
LEP	0.00	0.00	0.03	0.01	0.04	0.01	0.04	0.04
Retention Status								
Not Retained	0.01	0.01	0.05	0.02	0.04	0.02	0.03	0.02
Retained	0.01	0.02	0.06	0.01	0.06	0.04	0.06	0.07
Special Education Status								
Not in SpEd	0.01	0.01	0.05	0.02	0.04	0.02	0.03	0.03
In SpEd	0.00	0.01	0.05	0.02	0.04	0.02	0.04	0.02
Overage Status								
Not Overage	0.00	0.00	0.04	0.01	0.03	0.01	0.03	0.02
Overage	0.01	0.02	0.07	0.04	0.06	0.03	0.04	0.05

Note. A "-" indicates that no maximum likelihood estimates were calculated for these analyses due to a quasi-complete separation in the data points.

[a] Analysis for this predictor included only time periods 10(1) through 12(2).

Across all time periods except for 11(1), it can be seen that males are always at a greater risk of withdrawing than females. The risk is especially high during time period 10(1), which is also the period of highest risk for females. Across each time period, Anglo students have the highest hazard probabilities, while African American students have the smallest probabilities for withdrawal. Each ethnic group has higher risks for withdrawal during the first semesters of each year across all four years. Surprisingly, students who are English Proficient (EP) are at a higher risk of withdrawing in most time periods than the LEP students. In all time periods except 10(2), students who have been retained are more likely to withdraw than those who have not, especially at time periods 10(1), 11(1), 12(1), and 12(2). There is not much difference in the risks of withdrawing for students who are or are not in special education programs. Students who are overage respective to their classmates have a much higher probability than their classmates of withdrawing across all time periods, particularly at time periods 10(1), 11(1), and 12(2). Perhaps these time periods, when the new school year starts or when it is almost time to graduate, are especially sensitive for the overage student.

Dropout

It is also the appropriate to use the model including the interaction with time to interpret the outcome of dropping out. (See Table 4 and Table 5 for the -2LL values.) No estimators were produced for Asian students in periods 12(1) and 12(2) because these students withdrew only in time periods 9(1) through 11(2). The hazard probabilities of dropping out in each time period for each predictor variable group are shown in Table 7.

Table 7

Hazard Probabilities of Dropping Out in Each Time Period
by Predictor Variables

	Time Periods							
Variable	9(1)	9(2)	10(1)	10(2)	11(1)	11(2)	12(1)	12(2)
Gender								
Male	0.04	0.09	0.03	0.05	0.02	0.04	0.05	0.03
Female	0.04	0.07	0.02	0.03	0.02	0.03	0.04	0.02
Ethnicity								
Anglo	0.06	0.09	0.01	0.02	0.01	0.01	0.02	0.01
Afro Am	0.03	0.04	0.01	0.02	0.01	0.02	0.05	0.03
Hispanic	0.05	0.09	0.01	0.03	0.02	0.03	0.04	0.03
Asian[a]	0.04	0.09	0.04	0.02	0.02	0.04	-	-
LEP Status								
EP	0.04	0.06	0.01	0.02	0.01	0.02	0.04	0.02
LEP	0.05	0.09	0.01	0.04	0.03	0.05	0.05	0.02
Retention Status								
Not Retained	0.04	0.06	0.01	0.02	0.01	0.02	0.04	0.02
Retained	0.04	0.10	0.04	0.08	0.05	0.07	0.07	0.05
Special Education Status								
Not in SpEd	0.04	0.06	0.01	0.02	0.01	0.02	0.04	0.02
In SpEd	0.05	0.09	0.02	0.05	0.01	0.05	0.04	0.02
Overage Status								
Not Overage	0.02	0.03	0.01	0.01	0.01	0.02	0.03	0.02
Overage	0.08	0.14	0.03	0.06	0.03	0.05	0.06	0.03

Note. A "-" indicates that no maximum likelihood estimates were calculated for these analyses due to a quasi-complete separation in the data points.

[a] Analysis for this predictor included only time periods 9(1) through 11(2).

Similar to the outcome of withdrawal, males are at a greater risk than females, particularly at time periods 9(2) and 10(2), although females also experience a high risk (7% chance) at time period 9(2). As with other predictors, all ethnic groups are at the greatest risk of dropping out in the 9th grade, particularly Hispanic, Asian, and Anglo students. For the next four time periods, [10(1) - 11(2)], Asian students maintain the greatest probability for dropping out. During most time periods, LEP students are more likely to drop out than EP students. Time period 9(2) has the highest risk for these students, a 9% probability of dropping out.

Students who have been retained maintain a higher risk than those who have not, with their highest risk periods in 9(2), 10(2), 11(2), and 12(1). Perhaps as these students approach the end of a school year, facing the possibility of being retained once again, they choose to drop out rather than experience the failure. Special education students exhibit a similar pattern to retainees, having a high risk of dropping out in time periods 9(2), 10(2), and 11(2). Students who are overage have the highest risks of dropping out than any other subgroup in this study. Their hazard probabilities are consistently higher than those who are not overage. The 9th grade seems to be the most difficult time for these students, with hazards of 8% for time period 9(1) and a whopping 14% for time period 9(2).

Graduation

Graduation could only occur during the last three time periods, consequently, interpretation of the main effect model is the most appropriate. The hazards for the probability of graduating during these time periods are in Table 8.

Table 8

Hazard Probabilities of Graduating in Each Time Period
by Predictor Variables

	Time Periods							
Variable	9(1)	9(2)	10(1)	10(2)	11(1)	11(2)	12(1)	12(2)
			Gender					
Male	-	-	-	-	-	0.32	0.25	0.99
Female	-	-	-	-	-	0.77	0.71	0.99
			Ethnicity					
Anglo	-	-	-	-	-	0.06	0.04	0.98
Afro Am	-	-	-	-	-	0.10	0.07	0.99
Hispanic	-	-	-	-	-	0.07	0.05	0.99
Asian	-	-	-	-	-	0.05	0.04	0.97
			LEP Status					
EP	-	-	-	-	-	0.01	0.01	0.85
LEP	-	-	-	-	-	0.05	0.04	0.98
			Retention Status					
Not Retained	-	-	-	-	-	0.01	0.01	0.85
Retained	-	-	-	-	-	0.05	0.04	0.98
			Special Education Status					
Not in SpEd	-	-	-	-	-	0.01	0.01	0.84
In SpEd	-	-	-	-	-	0.05	0.04	0.98
			Overage Status					
Not Overage	-	-	-	-	-	0.01	0.01	0.94
Overage	-	-	-	-	-	0.07	0.05	0.99

Note. Analyses for this outcome included only time periods 11(2) through 12(2).

Females have a much higher probability of early graduation in time periods 11(2) and 12(1) than males. However, if males remain in the risk pool until time period 12(2), they have the same chance (99%) of graduating as the female students. A surprising result for the ethnic predictors is that, for this data set, both African American and Hispanic students have a higher probability of graduating in each time period than the Anglo students. African American students have a 10% probability of graduating at the end of the 11th grade and a 7% chance after first semester of the 12th grade; much higher probabilities than any other ethnic group. But for all ethnic groups, if students remain through time period 12(2), they have very high probabilities of graduating. Another interesting finding is that the LEP students have consistently higher probabilities of graduating than the EP students. EP students, who make up 91.6% of the data set, have only an 85% probability of graduating, compared to the 98% chance of the LEP students.

Students who have been retained and students who are enrolled in special education have equal probabilities of graduating in each time period. Again, their chances are higher than that of their counterparts', those who have not been retained and are not in special education. If these students remain in school through time period 12(2), they have a 98% probability of graduating. Overage students who have not experienced some other mode of exit before the 12th grade also have a high probability (99%) of graduating.

Still Enrolled After Four Years of High School

The outcome of being still enrolled after four years of high school has only one appropriate time period to predict, that of 12(2), therefore, the main effect model is used. Hazard probabilities for each predictor variable for time period 12(2) are listed in Table 9.

Table 9

Hazard Probabilities of Being Still Enrolled in Each Time Period by Predictor Variables

	Time Periods							
Variable	9(1)	9(2)	10(1)	10(2)	11(1)	11(2)	12(1)	12(2)
				Gender				
Male	-	-	-	-	-	-	-	0.98
Female	-	-	-	-	-	-	-	0.99
				Ethnicity				
Anglo	-	-	-	-	-	-	-	0.57
Afro Am	-	-	-	-	-	-	-	0.74
Hispanic	-	-	-	-	-	-	-	0.62
Asian	-	-	-	-	-	-	-	0.53
				LEP Status				
EP	-	-	-	-	-	-	-	0.14
LEP	-	-	-	-	-	-	-	0.55
				Retention Status				
Not Retained	-	-	-	-	-	-	-	0.14
Retained	-	-	-	-	-	-	-	0.55
				Special Education Status				
Not in SpEd	-	-	-	-	-	-	-	0.04
In SpEd	-	-	-	-	-	-	-	0.54
				Overage Status				
Not Overage	-	-	-	-	-	-	-	0.17
Overage	-	-	-	-	-	-	-	0.60

Note. Analyses for this outcome included only time period 12(2).

Male and female students are equally likely to be still enrolled. When comparing ethnic groups, African American and Hispanic students are more likely than Anglo (57%) or Asian students (53%), to remain longer than four years, presumably because the latter groups have experienced some other outcome in previous time periods. Those that continue in school after four years are also equally likely (54%) to be LEP, in special education, and have been retained in previous years. Students who are overage have a 60% probability of this outcome, the highest of all predictor groups.

No-Show

There were no students on the database coded as no-shows during the 9th grade year, probably due to data input error rather than actual nonoccurrence. The main effect model produced the most change in the -2LL. The hazard probabilities for this outcome are presented in Table 10.

Table 10

Hazard Probabilities of Being a No-Show in Each Time Period by Predictor Variables

	Time Periods							
Variable	9(1)	9(2)	10(1)	10(2)	11(1)	11(2)	12(1)	12(2)
			Gender					
Male	-	-	0.10	0.26	0.38	0.50	0.78	0.22
Female	-	-	0.44	0.72	0.82	0.88	0.96	0.67
			Ethnicity					
Anglo	-	-	0.00	0.01	0.02	0.03	0.11	0.00
Afro Am	-	-	0.01	0.02	0.04	0.06	0.19	0.02
Hispanic	-	-	0.00	0.02	0.03	0.01	0.13	0.01
Asian	-	-	0.00	0.01	0.02	0.03	0.10	0.01
			LEP Status					
EP	-	-	0.00	0.00	0.00	0.00	0.02	0.00
LEP	-	-	0.00	0.01	0.02	0.03	0.10	0.01
			Retention Status					
Not Retained	-	-	0.00	0.00	0.00	0.00	0.01	0.00
Retained	-	-	0.00	0.01	0.02	0.03	0.11	0.01
			Special Education Status					
Not in SpEd	-	-	0.00	0.00	0.00	0.00	0.02	0.00
In SpEd	-	-	0.00	0.01	0.02	0.03	0.10	0.01
			Overage Status					
Not Overage	-	-	0.00	0.00	0.00	0.01	0.02	0.00
Overage	-	-	0.01	0.02	0.03	0.04	0.13	0.01

Note. Analyses for this outcome included only time periods 10(1) through 12(2).

In every time period after the 9th grade, females had higher probabilities of being no-shows than males. Their hazards are alarmingly high, especially in time periods 10(2) through 12(1), where probabilities ranged from 72% to 96%. Likewise, time period 12(1) seems to be a critical period for all predictor groups, with highest probabilities occurring during this period. If students are likely to be no-shows, this seems to happen most frequently at the beginning of time period 12(1).

No Known Outcome

There is an increased need for accountability regarding students' mode of exit from school. Whether there are errors in data input or it is truly not known what becomes of certain students, schools must make every possible effort to ascertain and correctly identify student outcomes. With this knowledge, schools can become aware of which outcomes are most probable for groups of students, and support efforts to keep students in school. In this data set, the mode of exit from high school for 7% of the students was not known. As with the model for no-shows, there were no unknown outcomes for students during the 9th grade. The main effect model is appropriate, therefore, because there are no hazards for time periods 9(1) and 9(2). Hazard probabilities for time periods 10(1) through 12(2) are listed in Table 11.

Table 11

Hazard Probabilities of Having No Known Outcome in Each Time Period by Predictor Variables

	Time Periods							
Variable	9(1)	9(2)	10(1)	10(2)	11(1)	11(2)	12(1)	12(2)
Gender								
Male	-	-	0.38	0.50	0.48	0.54	0.75	0.53
Female	-	-	0.82	0.88	0.87	0.89	0.96	0.89
Ethnicity								
Anglo	-	-	0.07	0.10	0.10	0.12	0.25	0.11
Afro Am	-	-	0.11	0.17	0.16	0.19	0.39	0.18
Hispanic	-	-	0.08	0.12	0.11	0.14	0.29	0.13
Asian	-	-	0.06	0.09	0.08	0.10	0.23	0.10
LEP Status								
EP	-	-	0.01	0.01	0.01	0.02	0.04	0.02
LEP	-	-	0.06	0.10	0.10	0.11	0.24	0.10
Retention Status								
Not Retained	-	-	0.01	0.02	0.01	0.02	0.04	0.01
Retained	-	-	0.07	0.10	0.10	0.11	0.24	0.10
Special Education Status								
Not in SpEd	-	-	0.01	0.01	0.01	0.02	0.04	0.01
In SpEd	-	-	0.06	0.10	0.09	0.11	0.23	0.10
Overage Status								
Not Overage	-	-	0.00	0.02	0.02	0.02	0.05	0.02
Overage	-	-	0.08	0.12	0.11	0.13	0.28	0.12

Note. Analyses for this outcome included only time periods 10(1) through 12(2).

Female students have very high probabilities of having no known outcome in each time period, much higher than male students. African American students have probabilities of having no known outcome that range from 11% in time period 10(1) to 39% in time period 12(1). These hazards are consistently higher than for any of the other ethnic groups, although hazards for Anglo and Hispanic are higher than for Asian. LEP students, special education students, and retainees have relatively equal hazards, always higher than their EP, regular education, non-retainee counterparts. Overage students have the next highest hazards of having no known outcome after the female and ethnic groups previously mentioned.

The last three time periods continue to remain critical for all predictor groups, with time period 12(1) having the highest hazards. This data set appears to "lose" more students during the fourth year than at any other time.

Same Results — Different Point of View

Ultimately, a competing risks survival analysis with this type of data should allow schools or school districts to ascertain which periods of time present the highest risk for different modes of exit from school for students with certain characteristics. Therefore, it is also useful to look at the various competing outcomes from the viewpoint of the predictor variables. Graphs are an effective way to demonstrate the power of this method. The hazards for dropping out during each time period for the four ethnic groups are graphed in Figure 1. A glance at any time period gives one a visual cue as to which ethnic group is at the greatest risk of dropping out or which time periods are riskiest for one or all groups.

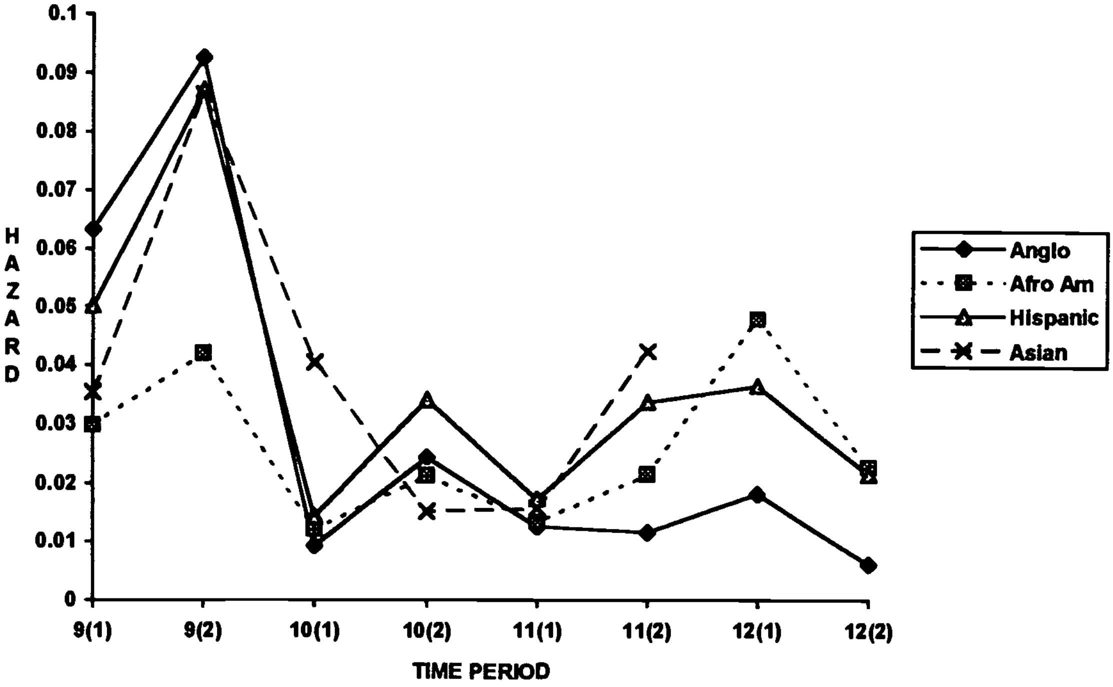

Figure 1. Hazard probabilities for dropping out for each time period by ethnic groups.

Combined graphs of each competing risk for a predictor variable can likewise show which risk is most likely during which time period. In Figure 2, the hazards for each competing risk are plotted for students who have been retained. The outcome of being still enrolled after four years occurs only after the last time period, 12(2), and has a "o" representing its hazard probability.

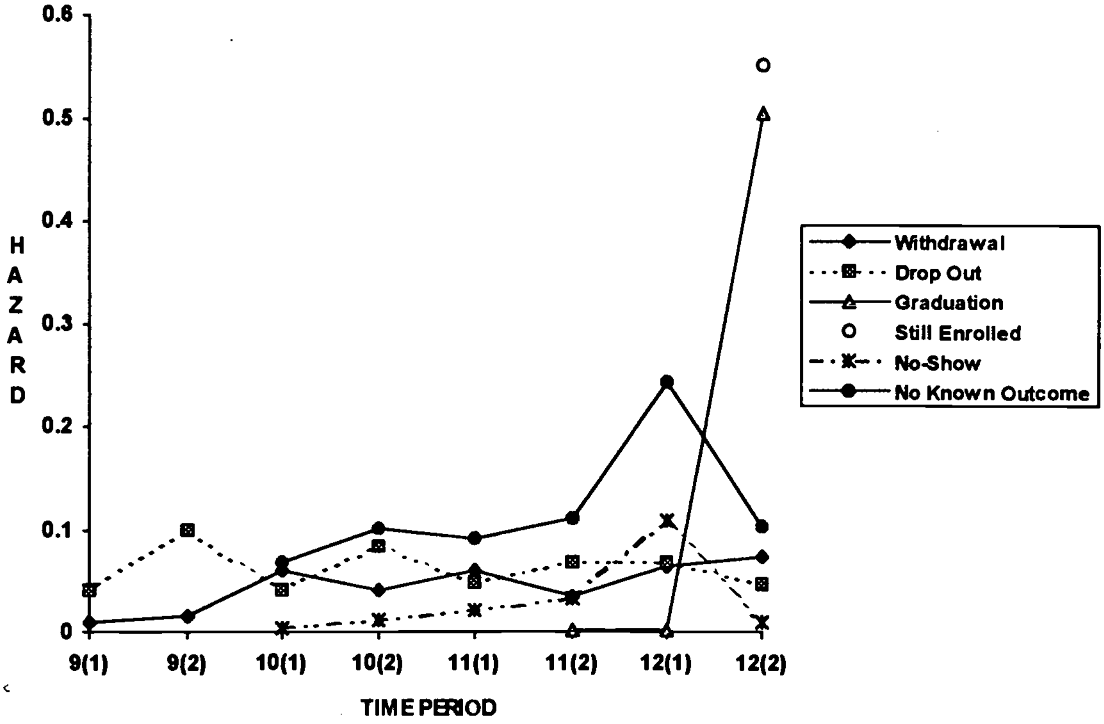

Figure 2. Hazards of each competing risk for students who have been retained.

An examination of this graph allows one to pinpoint the most and least likely outcomes for a student who has been retained during each time period or which outcomes are more likely across all time periods. For students who have been retained, the highest probability across time periods 10(1) through 12(1) is to have no known outcome. This kind of information should allow schools to see that this particular group of students has a tendency to "get lost in the crowd" and may require special attention.

Conclusions

Although there are a variety of statistics available to describe individual modes of exit from high school, only the survival analysis model using logistic regression computes the *probabilities* of the occurrence of the event in question. Consider the following example of three different statistics that can be calculated regarding student dropout: (a) the percentage of students remaining in school or choosing some other method of exit,

(b) the percentage of all dropouts who left during each time period, and (c) the hazard probabilities, or risk, of dropping out in each time period. A comparison of three different computational approaches is presented in Table 12, using information from the Dallas Public Schools data set.

Table 12

A Comparison of Three Different Computational Approaches

Time Periods	Percentage remaining in school (% surviving)	Percentage of all dropouts leaving each time period	Hazard probabilities of dropping out
9(1)	95.8	20.8	4.2
9(2)	89.5	30.7	6.6
10(1)	88.3	5.6	1.3
10(2)	86.2	10.3	2.6
11(1)	85.1	5.6	1.5
11(2)	83.4	8.5	2.4
12(1)	80.8	12.9	3.9
12(2)	79.6	5.5	1.9

Note. Adapted from Willett and Singer, 1991, Table 2.

Examination of the second column in Table 12 reveals the percentages of students from the initial data set that either remained in school or chose some other method of exiting for each time period. There is a steadily declining number of students that "survive" dropping out. By reviewing the third column, one could conclude that periods 9(1), 9(2), 10(2), and 12(1) are times when most students are likely to drop out. The hazard probabilities for dropping out, by time period, are presented in the fourth column. Examination of the hazard probabilities by time periods leads one to several different conclusions. Because the hazard probability modifies the risk set — earlier dropouts are omitted from the analysis because they are no longer eligible to drop out (having already done so) — it is evident that periods 9(1), 9(2), and 12(1) are critical. However, for periods 10(1) and 12(2) the orders are reversed, period 10(1) being actually more "risky" than period 12(2).

Willett and Singer (1991) discuss the

> apparent anomaly that arises from the differences in the definitions of risk reflected in the two summary statistics. Both sets of summary statistics are 'correct,' they simply answer different questions. Both identify periods of high risk, but they refer to different groups of students. Column 3 answers the question: For a randomly selected *high-school dropout*, when is dropout most likely to occur? Column 4 answers the question: For a randomly selected *currently enrolled high school student*, when is dropout most likely to occur? Although examining the proportion of dropouts who leave in each time period characterizes the population of dropouts, it does not describe the risk of dropping out over time among students in school (p. 434).

Extending the above example to the use of the competing risks survival analysis model, it can now be asked "For a randomly selected currently enrolled high school student, *what outcome is most likely to occur, and during which time period*?" The use of various predictor variables further extends the question to: "For a randomly selected currently enrolled *female* high school student, what outcome is most likely to occur, and during which time period?" or "For a randomly selected currently enrolled *African American* high school student, what outcome is most likely to occur, and during which time period?"

The predictor variables used for the competing risks survival analysis in the present paper are only a few of the many that could have been used. Other interactions that might have been incorporated include the cross-products between gender and the ethnic predictors, overage status, special education status, or retention status and these cross-products interaction with time. Although Hachen (1988) and Singer and Willett (1993) warn against including too many variables in the logistic regression equation, cross-products with gender and other predictors could produce informative prediction probabilities. Family and school characteristics are other sources of information that might prove to be significant.

The competing risks survival analysis method has received little use thus far in the field of education, but it merits a closer look as schools make an effort to educate a more diverse group of students who are faced with more choices than graduation or dropping out. A more precise prediction of the probability of these various modes of exiting school can allow decision-makers to initiate various remediation or intervention programs designed to keep students in school until graduation. Knowledge of the appropriate timing for these programs is essential in terms of the cost of development and human resources needed for successful programs. "Survival methods offer educational researchers much more than just a sophisticated data analytic approach — they offer a unified framework for appropriately modeling the many paths real students take throughout real schools" (Willett and Singer, 1993, p. 427).

References

Agresti, A. & Finlay, B. (1988). *Statistical methods for the social sciences*. San Francisco: Dellen Publishing.

Allison, Paul D. (1984). *Event history analysis: Regression for longitudinal event data*. Sage University Paper Series on Quantitative Applications in the Social Sciences, Series No. 07-046. Beverly Hills, CA: Sage Publications.

Cox, David (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B*, *34*, 187-202.

Denson, K. & Schumacker, R.E. (1994, April). *A Johnson-Neyman-like approach to interpreting significant discrete-time periods in survival analysis*. Paper presented at the annual American Educational Research Association meeting, New Orleans, LA.

Ensminger, M.C. & Slusarcick, A.L. (1992). Paths to high school graduation or dropout: A longitudinal study of a first grade cohort. *Sociology of Education*, *85*(2), 95-113.

Ferguson, G.A. & Takane, Y. (1989). *Statistical analysis in psychology and education*. New York: McGraw-Hill Book Company.

Fitzpatrick, K.M. & Yoels, W.C. (1992). Policy, school structure, and sociodemographic effects on statewide high school dropout rates. *Sociology of Education*, *65*(1), 76-93.

Hachen, David (1988). The competing risks model: A method for analyzing processes with multiple types of events. *Sociological Methods and Research*, *17*(1), 21-54.

Hamilton, Lawrence (1992). *Regression with graphics: A second course in applied statistics*. Belmont, CA: Duxbury Press.

Hildebrand, David (1986). *Statistical thinking for behavioral scientists*. Boston, MA: PWS Publishers:.

Kortering, L.J. & Blackerby, J. (1992). High school dropout and students identified with behavioral disorders. *Behavioral Disorders*, *18*(1), 24-32.

Lakebrink, J.M. (1989). A gender at risk. In J.M. Lakebrink (Ed.), *Children at risk*. Springfield, IL: Charles C. Thomas, 216-229.

Miller, A.P. (1989). Student characteristics and the persistence/dropout behavior of Hispanic students. In J.M. Lakebrink (Ed.), *Children at risk*. Springfield, IL: Charles C. Thomas, 119-139.

McMillen, M.M. (1994). *Dropout rates in the United States*. Washington, DC: National Center for Education Statistics (ED)

Menard, Scott (1995). *Applied logistic regression analysis*. Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-106. Thousand Oaks, CA: Sage.

Morrow, George (1986). Standardizing practice in the analysis of school dropouts. *Teachers College Record, 87*(3), 342-355.

Murnane, R.J., Singer, J.D. & Willett, J.B. (1988). The career paths of teachers: Implications for teacher supply and methodological lessons for research. *Educational Researcher, 17*(6), 22-30.

Murnane, R.J., Singer, J.D. & Willett, J.B. (1989). The influences of salaries and "opportunity costs" on teachers' career choices: Evidence from North Carolina. *Harvard Educational Review, 59*(3), 325-346.

Nason, R.B. (1991). Retaining children: Is it the right decision? *Childhood Education, 67*(5), 300-304.

Neter, J., Wasserman, W. & Kutner, M.H. (1989). *Applied linear regression models*. Homewood, IL.: Irwin Publishing.

Orr, M.T. (1987). *Keeping students in school: A guide to effective dropout prevention programs and services*. San Francisco, CA: Jossey-Bass.

Pittman, R.B. (1995). The potential high school dropout, the 21st century, and what's ahead for rural teachers. *Rural Educator, 16*(3), 23-27.

Roderick, Melissa (1994). Grade retention and school dropout: Investigating the association. *American Educational Research Journal, 31*(4), 729-759.

Rumberger, R.W. (1995). Dropping out of middle school: A multilevel analysis of students and schools. *American Educational Research Journal, 32*(3), 583-625.

Singer, J.D., Fosburg, S., Goodson, B.D., & Smith, J.M. (1978). National Day Care Home Study Research Report. Final Report of the National Day Care Home Study. DHHS Publication No. 80-30283.

Singer, J. D. & Willett, J. B. (1991). Modeling the days of our lives: Using survival analysis when designing and analyzing longitudinal studies of duration and the timing of events. *Psychological Bulletin, 110*(2), 268-290.

Singer, J. D. & Willett, J. B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Research*, 18(2), 155-195.

Sween, J.A. (1989). The timing of dropping out, the possibility of early intervention, and the need for intervention before high school. In J.M. Lakebrink (Ed.), *Children at risk*. Springfield, IL: Charles C. Thomas, 32-44.

Watt, D. & Roessingh, H. (1994). ESL dropout: The myth of educational equity. *Alberta Journal of Educational Research, 40*(3), 283-296.

Willett, J. B. & Singer, J. D. (1991). From whether to when: New methods for studying student dropout and teacher attrition. *Review of Educational Research, 61*(4), 407-450.

Willett, J. B. & Singer, J. D. (1991). How long did it take . . .?: Using survival analysis in psychological research. In L. M. Collins & J.L. Horn (Eds.), *Best methods for the analysis of change: Recent advances, unanswered questions, future directions.* Washington, DC: American Psychological Association, 309-326.

Wright, Raymond (1995). Logistic regression. In L.C. Grimm & P.R. Yarnold (Eds.) *Reading and understanding multivariate statistics.* Washington, DC: American Psychological Association, 217-244.

Zwick, R. & Braun, H.I. (1988). *Methods for analyzing the attainment of graduate school milestones: A case study* (GRE Board Professional Report No. 86-3P; ETS Research Report No. 88-30). Princeton, NJ: Educational Testing Service.

Appendix A

WILLETT AND SINGER'S SAS PROGRAM, MODIFIED TO CONDUCT A COMPETING RISKS SURVIVAL ANALYSIS

```
* CREATING THE PERSON-PERIOD DATA SET;

DATA ALL;
        SET COMPRISK; (Assumes the previous creation of dataset COMPRISK)
        ARRAY OCCASION[8]E1-E8;
* TO CREATE GENDER * TIME;
        ARRAY SEXTIME[8]SX1-SX8; (Creates the variable SEXTIME)
* TO CREATE ANGLO * TIME;
        ARRAY ANGTIME[8]A1-A8; (Creates the variable ANGTIME)
* TO CREATE LEP * TIME;
        ARRAY LEPTIME[8]L1-L8; (Creates the variable LEPTIME)
```

(Continue until all predictor variables have been crossed with time)

```
        DO PERIOD=1 TO MIN(LASTPD8,8);
                IF PERIOD=LASTPD8 AND WD=1 THEN Y=1;
                        ELSE Y=0;
```

(Change WD to DROP, GRAD, STILLIN, NOSHOW, NOKNOW for other outcomes)

```
        DO INDEX=1 TO 8;
                IF INDEX=PERIOD THEN OCCASION[INDEX]=1;
                        ELSE OCCASION[INDEX]=0;
                SEXTIME[INDEX]=SEX*OCCASION[INDEX];
        END;
        DO INDEX=1 TO 8;
                IF INDEX=PERIOD THEN OCCASION[INDEX]=1;
                        ELSE OCCASION[INDEX]=0;
                ANGTIME[INDEX]=ANGLO*OCCASION[INDEX];
        END;
        DO INDEX=1 TO 8;
                IF INDEX=PERIOD THEN OCCASION[INDEX]=1;
                        ELSE OCCASION[INDEX]=0;
                LEPTIME[INDEX]=LEP*OCCASION[INDEX];
        END;
```

(Continue until all predictor variables have been crossed with OCCASION)

```
END;
OUTPUT;
END;

*CREATING THE INITIAL MODEL;

PROC LOGISTIC DATA=ALL NOSIMPLE OUT=ESTIMATE DESCENDING;
TITLE2 "INITIAL (NULL) MODEL";
MODEL Y=E1-E8/NOINT MAXITER=20;
```

```
*COMPUTING FITTED HAZARD AND SURVIVAL FUNCTIONS;

DATA NEWEST;
        SET ESTIMATE;
        ARRAY OCCASION[8]E1-E8;
        SURVIVAL=1;
        DO PERIOD=1 TO 8;
                X=OCCASION[PERIOD]
                HAZARD=1/(1+(EXP(X)));
                SURVIVAL=(1-HAZARD)*SURVIVAL;
                OUTPUT;
        END;

*PRINT SURVIVAL AND HAZARD RESULTS;

PROC PRINT;
        VAR PERIOD SURVIVAL HAZARD;
        FORMAT SURVIVAL HAZARD 6.4;
PROC PLOT;
        PLOT(SURVIVAL HAZARD)*PERIOD;

*MODEL WITH MAIN EFFECT OF GENDER;

PROC LOGISTIC DATA=ALL NOSIMPLE OUT=ESTIMATE DESCENDING;
TITLE2 "MAIN EFFECT OF GENDER";
MODEL Y=E1-E8  SEX/NOINT MAXITER=20;

*COMPUTING FITTED HAZARD AND SURVIVAL FUNCTIONS;

DATA NEWEST;
        SET ESTIMATE;
        ARRAY OCCASION[8]E1-E8;
        DO SEX=1 TO 2;
                SURVIVAL=1;
                DO PERIOD=1 TO 8;
                        X=OCCASION[PERIOD]+(SEX-1)*SEX;
                        HAZARD=1/(1+(EXP(X)));
                        SURVIVAL=(1-HAZARD)*SURVIVAL;
                OUTPUT;
                END;
        END;
```

```
*PRINT SURVIVAL AND HAZARD RESULTS;

PROC SORT;
        BY SEX;
PROC PRINT;
        BY SEX;
        ID PERIOD;
        VAR SURVIVAL HAZARD;
        FORMAT SURVIVAL HAZARD 6.4;
PROC PLOT;
        PLOT(SURVIVAL HAZARD)*PERIOD=SEX;
```

(Continue until all main effect models have been created)

```
*MODEL WITH INTERACTION BETWEEN GENDER AND TIME;
*THESE MODELS TEST THE ASSUMPTION OF PROPORTIONALITY;

PROC LOGISTIC DATA=ALL NOSIMPLE OUT=ESTIMATE DESCENDING;
TITLE2 "INTERACTION BETWEEN GENDER AND TIME”;
MODEL Y=E1-E8 SX1-SX8/NOINT MAXITER=20;

*COMPUTING FITTED HAZARD AND SURVIVAL FUNCTIONS;

DATA NEWEST;
        SET ESTIMATE;
        ARRAY OCCASION[8]E1-E8;
        ARRAY SEXTIME[8]SX1-SX8;
        DO SEX=1 TO 2;
                SURVIVAL=1;
                DO PERIOD=1 TO 8;
                        X=OCCASION[PERIOD]+(SEX-1)*SEXTIME[PERIOD];
                        HAZARD=1/(1+(EXP(X)));
                        SURVIVAL=(1-HAZARD)*SURVIVAL;
                OUTPUT;
                END;
        END;

*PRINT SURVIVAL AND HAZARD RESULTS;

PROC SORT;
        BY SEX;
PROC PRINT;
        BY SEX;
        ID PERIOD;
        VAR SURVIVAL HAZARD;
        FORMAT SURVIVAL HAZARD 6.4;
PROC PLOT;
        PLOT(SURVIVAL HAZARD)*PERIOD=SEX;
```

(Continue until all interaction models have been created)

Appendix B

Withdrawal and Dropout Reasons

Withdrawal Reasons

Death
Institutionalization
In approved GED program
Job training center
Night school
Transfer to a private school
Transfer to another district

Dropout Reasons

Age
Dislike school
Employment
Expulsion
Low or failing grades
Marriage
Non-approved GED program
Non-permanent resident
Pregnancy
Socio-economic reasons
Transfer to another school with no documentation
Transfer to a non-approved program
30 consecutive absences

Appendix C

Transformation of Data from the Person-Data to the Person-Period Data Set

The following is an input line of Person-Data for an Anglo female student who withdrew during the third time period, 10(1).

The SAS program in Appendix A transforms the one line of data to three lines of data, one for each time period that the student was enrolled.

American Educational Research Association
April 18-22, 1995

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC®

REPRODUCTION RELEASE

(Specific Document)

TM025995

306
UD031~~305~~

AERA /ERIC Acquisitions
The Catholic University of America
210 O'Boyle Hall
Washington, DC 20064

I. DOCUMENT IDENTIFICATION:

Title: Student Choices: Using a Competing Risks Model of Survival Analysis	
Author(s): Katy Denson and Randall E. Schumacker	
Corporate Source:	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce the identified document, please CHECK ONE of the following options and sign the release below.

☒ Sample sticker to be affixed to document

Sample sticker to be affixed to document □

Check here

Permitting microfiche (4" x 6" film), paper copy, electronic, and optical media reproduction

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

____

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 1

"PERMISSION TO REPRODUCE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY

____

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Level 2

or here

Permitting reproduction in other than paper copy

Sign Here, Please

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at Level 1.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Signature: Kathleen Denson	Position: Program Evaluator
Printed Name: Kathleen Denson	Organization: Dallas Public Schools
Address: 3801 Herschel, Dallas TX 75204-5491	Telephone Number: ( ) (214) 599 5328 Date: July 10, 1996

You can send this form and your document to the ERIC Clearinghouse on Assessment and Evaluation. They will forward your materials to the appropriate ERIC Clearinghouse. ERIC/AERA Acquisitions, ERIC Clearinghouse on Assessment and Evaluation, 210 O'Boyle Hall, The Catholic University of America, Washington, DC 20064, (800) 464-3742