An Investigation on Language Programs in US Higher Institutions—A Case Study on Chinese Language Programs

Huiqiang Zheng
University of Iowa, Iowa, USA
Lukun Zheng
Western Kentucky University, Bowling Green, USA

Although there have been an increasing number of programs offered by universities in the United States (U.S.) in recent years, few published studies address the development of any in U.S. universities. Therefore, the current study aims to address important issues to the success of the language program and bridge the gap between the existing literature and language program development implementing statistical methods of ordinal logistic regression and factor analysis. Based on data collecting from foreign language programs in 201 different universities across the U.S., we developed a proportional odds model to examine the factors that foster or constrain the development of the program. The study can also be used as a predictive model on the perspective of starting a new language program in a university based on its features in different respects. An internal validation of our model was conducted using the cross-classification table.

Keywords: foreign language programs, ordinal logistic regression, proportional odds models, predicted classification

Introduction

There has been an increasing interest in learning Chinese in recent years, as can be seen in the increasing number of Chinese programs offered by universities in the United States (U.S.) (Nelley, Goldberg, & Lusin, 2010). Based on a survey conducted by the Modern Language Association, Welles (2004) reported a 20% increase in enrollments in university Chinese classes from 1998 to 2002. The number of Chinese programs in the U.S. for learners of all ages tripled from 1995 to 2005 and has continued to increase in last decades (Dobuzinskis, 2011; Neely, 2011; Rogers, 2012). This increase is mainly due to globalization and governmental interests, especially considering China's growing economic and cultural importance in the modern world. For instance, in September 2005, the National Security Education Program at the Department of Defense, funded the University of Oregon and Portland Public Schools for a program to develop a national model in graduating students who are culturally and linguistically competent in Chinese (OELA Newsline, 2005). Modern information and communication technologies (ICT) has gradually been applied in Chinese teaching and learning, as computers and the Internet, being free of time and location constraints, have been standard facilities in almost all U.S. universities. Studies indicate that the ICT helps create or foster

---

Huiqiang Zheng, Ph.D., lecturer, Department of Asian and Slavic Languages and Literatures, University of Iowa.
Lukun Zheng, Ph.D., assistant professor, Department of Mathematics, Western Kentucky University.

classroom-based language-and-culture curricula (Grant & Huang, 2010), blended-learning environments (Huang, Lin, & Chiang, 2010), knowledge sharing (Lin & Wang, 2008), collaborative learning tasks (K. Chang, Lan, C. Chang, & Sung, 2010), and language acquisition (Chen & Liu, 2008). This modern technology integration into teaching and learning Chinese has helped to expand the instruction options, curriculum, and faculty development, which has brought more potential to the development of Chinese programs.

The flourish of interests and activities has resulted in an increased emphasis on the possibility to start new Chinese programs in U.S. universities. However, due to lack of experience and other reasons, some newly started Chinese language programs turn out to be unsuccessful, with low enrollments and few Chinese courses to offer. Some Chinese programs even got shut down after running for several years. Hence, it is our concern to study the related factors in the development of Chinese programs in U.S. universities. The success of Chinese programs in universities in the U.S. relies on many factors, like financial budget, university location, ranking, and so on. The issue about assessing factors affecting the success of Chinese programs for U.S. universities is important, and yet little research has examined this topic.

This study is significant for many reasons. First, the interest of Chinese learning has been growing fast over the last two or three decades. The gap between the existing literature on Chinese program development and issues which are important to the development of Chinese programs makes this empirical study valuable. Second, in addition to descriptive statistical summaries, we adopted an inferential statistical approach to investigate the significant factors related to the success of Chinese programs using a generalized logistic regression model. The study is based on a large sample of Chinese programs at different levels in U.S. universities of different regions. Third, this study examines the significant factors for the success of Chinese programs. It is aimed to help the Chinese program directors and teachers as they do program evaluation and planning. Finally, the generalized logistic model developed in this study can also be used as a predictive model to help universities, which have not started Chinese programs to begin doing so by providing the predicted probabilities of different Chinese program status.

The remainder of this paper is organized as follows. Section two gives a description of experimental framework used in this study including a brief introduction of the generalized logistic regression method. Section three is devoted to the findings of this research in which a descriptive summary of the data is given together with a thorough statistical analysis of the model implied in this study. In Section four, we will draw our conclusion and discuss related future works.

Methodology

In this study, we collected data from a random sample of 201 universities and colleges in the U.S., which offer Chinese courses. The variables included in our data set include one response variable and nine explanatory variables. We give detailed summaries about the data set using some descriptive statistical methods. Also, we apply generalized logistic model to analyze the data collected in order to study the causal, interactive, and (or) moderating relationships among the explanatory variables and the response variable.

The Data Collection

Based on the results generated by the official websites providing applicants with relevant information about degree programs, Chinese programs are offered by numerous American Higher Education Institutions, namely, Stanford University, Yale University, University of Pennsylvania, Duke University, University of

Michigan at Ann Arbor, Brown University, University of Kentucky, and others (Graphiq, 2017). According to Graphiq (2007), there are 948 higher education institutions which offer Chinese programs. We drew a random sample of 276 higher institutions using simple random sampling method. However, among these 276 higher institutions, there are 75 of them which do not offer any Chinese programs based on our investigation. The study is based on the remaining 201 higher institutions which offer Chinese programs.

The response variable (y) in our study is the Chinese program status with three different categories. It equals 1 if it offers Chinese courses but without any degree on Chinese; it equals 2 if it offers a Chinese minor; and it equals 3 if it offers a Chinese major or a graduate program in Chinese. Table 1 provides the frequencies of these three different levels.

Table 1

Chinese Program Status Categories and Frequencies (Relative Frequencies) for the Study Sample

Category	Description	Frequency (Relative frequency)
1	Chinese courses but no degrees	108 (53.7 %)
2	Chinese minor	43 (21.4 %)
3	Chinese major or above	50 (24.9 %)

Nine potential explanatory variables or predictors included in the study are university type (U_type) (0 if public; 1 if private), university enrollment (U_enroll), university acceptance rate (A_rate), university rank (U_rank) (1—national university; 2—regional university; and 3—others based on 2017 US News and World Report ranking on colleges) (U.S. News and World Report, 2017), in-state and out-of-state tuition fees (T_in and T_out) (for private universities, these tuition fees will be the same), the number of foreign language minors offered in the university (x_minors), the number of foreign language majors offered in the university (x_majors), and the number of foreign language graduate programs offered in the university (x_grads). The foreign languages include Arabic, American Sign Language, Chinese, French, German, Greek, Hebrew, Italian, Japanese, Korean, Latin, Portuguese, Russian, and Spanish. These 14 languages are the most commonly learned foreign languages in U.S. higher education institutions according to Fall 2013 enrollments (Goldberg, Looney, & Lusin, 2015).

Among these variables, U_enroll, A_rate, T_in, T_out, x_minors, x_majors, and x_grads are numeric. U_type and U_rank are categorical variables. For statistical regression purpose, we will create two dummy variables Rank_D1 and Rank_D2 for U_rank since it has three different levels as follows:

$$RankD1 = \begin{cases} 1, national\ univeristy \\ 0, Otherwise \end{cases}, \quad RankD2 = \begin{cases} 1, regional\ univeristy \\ 0, Otherwise \end{cases} \qquad (1)$$

These two dummy variables will represent U_rank in our regression model. Hence, in our regression model, there are in total 10 potential predictors.

We adopted some of these variables developed in Graphiq website (Graphiq, 2017) like university enrollment, university acceptance rate, and so on. We include other variables like the number of language minors and majors offered in the university based on the interest of our study. These selected variables are suitable for this study, because they are from standard survey especially constructed to evaluate language programs in higher education institutions (Norris, 2016).

The Proportional Odds Logistic Regression Model

The classical logistic regression model or logit model is a regression model when the response variable is a binary categorical variable. There are two different generalizations for the logistics regression model when

the response variable has more than one possible category. One is called "the multinomial logistic regression for nominal response variables." The other is called "the ordinal logistic regression for ordinal response variables." In this study, we chose ordinal logistic regression since the response in the study is ordinal.

There are several different types of logistic regression models, which are well known for ordinal response variables, including the proportional odds model, the continuation ratio model, the adjacent categories logistic regression model, and the stereotype logistic model. The proportional odds model is the most commonly used one among them (Agresti, 2002, 2010; Armstrong & Sloan, 1989; Clogg & Shihadeh, 1994; Hilbe, 2009; Hosmer & Lemeshow, 2013; Liu, 2009, 2014; O'Connell, 2006; O'Connell & Liu, 2011).

The proportional odds logistic regression model is used to examine the relationship between the ordinal response variable and a set of p explanatory variables or predictors, which can be categorical or numerical. It estimates the cumulative odds and the probability of being at or below a specific level of the response for an observation, given the values of the predictors. Let the response be $Y = 1, 2, \ldots, J$, where the ordering is natural. The associated probabilities are $\{\pi_1, \pi_2, \ldots, \pi_J\}$, and a cumulative probability of a response less than or equal to j is:

$$P(Y \le j) = \tau_1 + \pi_2 + \cdots + \pi_j \tag{2}$$

The cumulative logit is defined as:

$$logit[P(Y \le j | x_1, x_2, \ldots, x_p)] = ln\left(\frac{P(Y \le j)}{1 - P(Y \le j)}\right) = ln\left(\frac{\tau_1 + \pi_2 + \cdots + \pi_j}{\tau_{j+1} + \pi_{j+2} + \cdots + \pi_J}\right) \text{ for } j = 1, 2, \ldots, J - 1 \tag{3}$$

This measures how likely the response is to be in category j or below versus in a category higher than j.

The proportional odds logistic regression model can be expressed as follows:

$$logit[P(Y \le j | x_1, x_2, \ldots, x_p)] = \alpha_j + \left(-\beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_p x_p\right) \text{ for } j = 1, 2, \ldots, J - 1. \tag{4}$$

This model has $(J - 1)$ intercepts plus p slopes, for a total of $r + p - 1$ parameters to be estimated. Notice that intercepts α_j can differ, but that slope for each variable stays the same across different equations. It estimates the cumulative logits across $J - 1$ response categories. One can obtain the estimated cumulative probabilities being at or below the j the category by taking the inverse logit:

$$P(Y \le j | x_1, x_2, \ldots, x_p) = \frac{exp[\alpha_j + \left(-\beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_p x_p\right)]}{1 + exp[\alpha_j + \left(-\beta_1 x_1 - \beta_2 x_2 - \cdots - \beta_p x_p\right)]} \tag{5}$$

From these cumulative probabilities, one can obtain the estimated probability for being in a particular category j through the equation:

$$P(Y = j | x_1, x_2, \ldots, x_p) = P(Y \le j | x_1, x_2, \ldots, x_p) - P(Y \le j - 1 | x_1, x_2, \ldots, x_p). \tag{6}$$

Here, we have $P(Y \le J | x_1, x_2, \ldots, x_p) = 1$ and $P(Y \le 0 | x_1, x_2, \ldots, x_p) = 0$.

Variable Selection

An ordinal logistic regression model was fitted in which the cumulative logit was estimated in a single model using maximum likelihood techniques (Archer & Williams, 2012; Hosmer & Lemeshow, 2013). For model building, backward selection procedures applying a *P*-value less than 0.01 for the log-likelihood ratio test were used to select the predictors in the best fitting model based on Akaike information criterion (AIC).

The *glmnetcr* package was used in our *R* programming of the ordinal logistic regression model. In fitting the model, we can extract the covariates into an object *x* and the ordinal response into the object *y*.
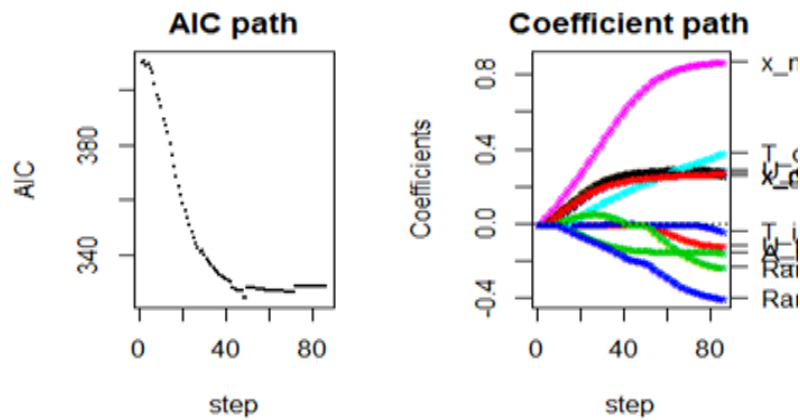
Figure 1. Plot of Akaike information criterion and estimated coefficients across the regularization path for the fitted glmnet.cr object.

The model attaining the minimum AIC is at Step 49, at which we extracted the best fitting model from the solution path. The predictors in the best fitting model includes: U_enroll, A_rate, T_out, x_minors, x_majors, x_grad, Rank_D1, and Rank_D2.

The resulting intercepts and coefficients can be used to calculate the predicted probabilities for each category as described in Section 2.2. For each observation, the predicted category is identified as the category with the highest predicted probability of all three categories. Classification tables can be obtained by cross-classifying the observed categories by the predicted categories. The model output and the classification tables are presented in the next section. The *R* code used to develop the ordinal logistic regression model is available on request. The analyses were performed with *R* version 3.4.0 (2017, The *R* Foundation for Statistical Computing, packages glmnetcr, oridnal, and MASS).

Table 2

Baseline Characteristics of 201 Randomly Selected Higher Institutions in Each of the Three Categories

Variables	Category 1 ($y = 1$)	Category 2 ($y = 2$)	Category 3 ($y = 3$)
University type			
Public (%)	49 (45.4)	26 (60.5)	22 (44)
Private (%)	59 (54.6)	17 (39.5)	28 (56)
University rank			
National (%)	32 (29.6)	21 (48.8)	35 (70.0)
Regional (%)	62 (57.4)	21 (48.8)	7 (14.0)
Others (%)	14 (13.0)	1 (2.4)	8 (16.0)
Enrollment mean (SD)	7,628.38 (7324.22)	17,023.23(10705.55)	16,425.44(14289.62)
Acceptance rate mean (SD)	66.91 (20.00)	66.91 (19.78)	49.70 (25.78)
T_out mean (SD)	27,202.33(11399.40)	26,767.12(9911.63)	35,925.54(12012.91)
x_minor mean (SD)	3.94(2.52)	6.86 (2.77)	9.38 (5.00)
x_major mean (SD)	3.22 (2.29)	3.77 (2.20)	7.18 (3.46)
x_grads mean (SD)	0.37 (0.88)	0.84 (1.27)	2.80 (3.64)

Results

Baseline Characteristics Information

The baseline characteristics for these 201 randomly selected higher institutions in each of the three categories (the categories of the response y) are presented in Table 2. We can see, for instance, most of the

higher institutions in Category 3 are national universities (70%). The average numbers of foreign language minors in Categories 1, 2, and 3 are 3.94, 6.86, and 9.38, respectively.

Figure 2 illustrates the distribution of quantitative variables selected using boxplots. For instance, from the boxplots for acceptance rate against different categories, we found that the distribution of acceptance rate for Categories 2 and 3 were about symmetric, while the distribution of acceptance rate for Category 1 was skewed to the left.
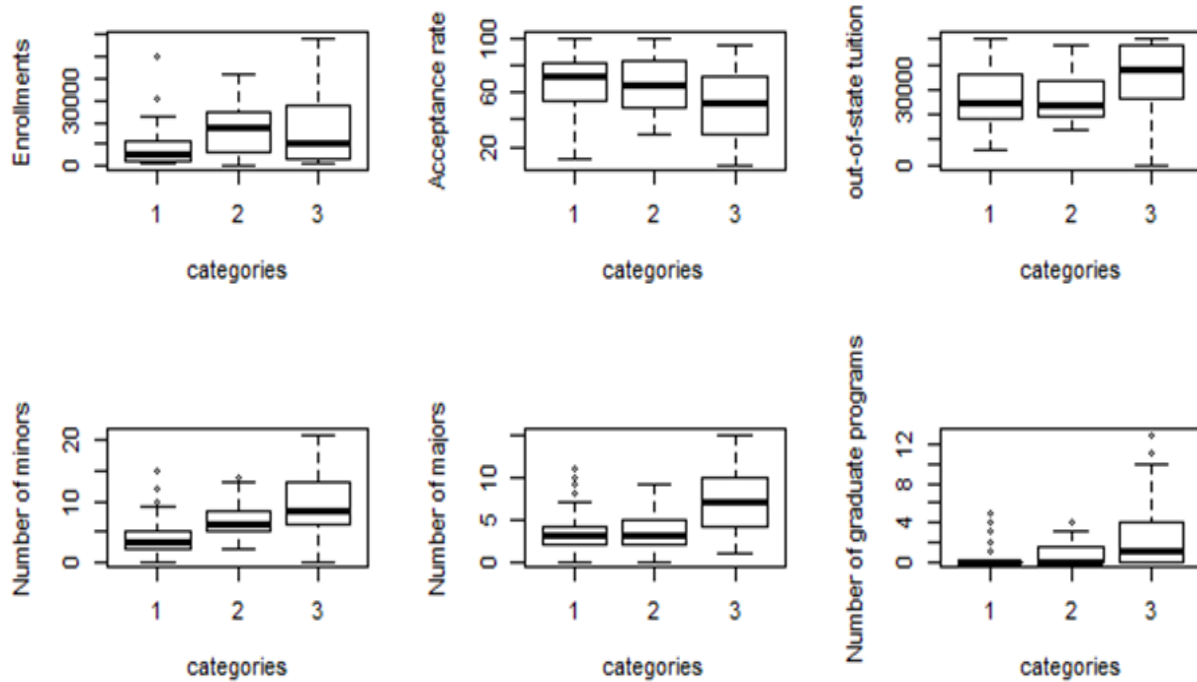

Figure 2. Boxplots of the quantitative variables: enrollment, acceptance rate, out-of-state tuition fee, x_minors, x_majors, and x_grads in each category.

Proportional Odds Model Analysis

The final proportional odds logistic regression model, in which the response variable *y* were modeled against the selected predictors U_enroll, A_rate, T_out , x_minors, x_majors, x_grad, Rank_D1, and Rank_D2 in the best fitting model, was fitted using *R*. Table 3 displays the *R* output for the model.

The first part of the output is the coefficient estimates for the slopes $\beta_1, \beta_2, \cdots, \beta_8$, and intercept estimates for the intercepts α_1 and α_2 in the proportional odds model. These are the estimates of the $J + p - 1 = 3 + 8 - 1 = 10$ unknown parameters in our model. The maximum likelihood estimates of the parameters are:

$$\hat{\beta}_1 = 3.221e - 05, \hat{\beta}_2 = -8.443e - 03, \hat{\beta}_3 = 2.861e - 05, \hat{\beta}_4 = 2.191e - 01,$$
$$\hat{\beta}_5 = 1.225e - 01, \hat{\beta}_6 = 1.289e - 01, \hat{\beta}_7 = -4.563e - 01, \hat{\beta}_8 = -9.131e - 01,$$
$$\alpha_1 = 2.062, \alpha_2 = 3.577$$

Based on the signs of the estimated parameters, U_enroll, T_out, x_minors, x_majors, and x_grad were positively associated with the log odds of being beyond a Chinese program status level. A_rate, Rank_D1, and Rank_D2 were negatively associated with the log odds of being beyond a Chinese program status level. In terms of odds ratio, for instance, the odds of being beyond a Chinese program status level were $exp(\hat{\beta}_4) = exp(0.2191) = 1.245$ times greater with a one unit increase in x_minors.

The last part of the output gives you the fit statistics where the "residual deviance" equals to $-2loglikelihood$ value for the current model. We can use this value to assess how well does this model fits in comparison to other models. For example, to compare it to the intercept only model, the log likelihood ratio Chi-Square test can be performed with the test statistic value being 103.6675 (the difference of the residual deviances between both models) and degree of freedom eight (difference of the number of parameters between both models). This is highly significant indicating that the fitted model is much better than the intercept only model.

Table 3

R Output of the Proportional Odds Model With the Selected Predictors by the Backward Selection Procedure

Formula:
y ~ U_enroll + A_rate + T_out + x_minors + x_majors + x_grad + Rank_D1 + Rank_D2

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
U_enroll	3.221e-05	2.064e-05	1.560	0.118717
A_rate	-8.443e-03	8.267e-03	-1.021	0.307107
T_out	2.861e-05	1.804e-05	1.586	0.112720
x_minors	2.191e-01	6.431e-02	3.406	0.000658
x_majors	1.225e-01	7.650e-02	1.602	0.109246
x_grad	1.289e-01	1.201e-01	1.073	0.283225
Rank_D1	-4.563e-01	5.697e-01	-0.801	0.423145
Rank_D2	-9.131e-01	5.551e-01	-1.645	0.100008

Threshold coefficients:

	Estimate	Std. Error	z value
1|2	2.062	1.101	1.872
2|3	3.577	1.125	3.179

Residual Deviance: 302.2552
AIC: 322.2552

Prediction

The resulting intercepts and coefficients can be used to calculate the predicted probabilities for each category as described in Section 2.2. For each observation, the predicted category is identified as the category with the highest predicted probability of all three categories. Classification tables can be obtained by cross-classifying the observed categories by the predicted categories as shown in Table 4.

Table 4

Cross-Classification Table of the Observed Categories by the Predicted Categories by Our Model

Observed categories		Category 1	Category 2	Category 3
	Category 1	96	29	16
Predicted categories	Category 2	7	5	4
	Category 3	5	9	30

From Table 4, we can see that the proportional odds model classified in total (96+5+4)/201 = 70.2% of all higher institutions in the study correctly.

The proportional odds model developed in this study can also be used as a predictive model to help universities which have not started Chinese programs to begin doing so. Given the values of the predictors

involved in our model, we can use our model to generate probabilities of potential Chinese program status categories. For instance, consider a higher institution with the following values:

U_enroll = 14,210, A_rate = 45.6, T_out = $30,120, x_minors = 6, x_majors = 3, x_grad = 0, Rank_D1 = 0, and Rank_D2 = 1.

That is, it is a regional university with an enrollment of 14,210, acceptance rate = 45.6, out-of-state tuition fee being $30,120, six foreign language minors, three foreign language majors, and no foreign language graduate programs.

Table 5

Predicted Probabilities for Different Categories for the Higher Institution With Given Attributes

	Category 1 ($y = 1$)	Category 2 ($y = 2$)	Category 3 ($y = 3$)
Predicted probabilities	0.8424410	0.1180785	0.0394805

Using *R*, we obtained the above probabilities for the higher institution with given attributes in Table 5. With this information, the university should consider offering Chinese course for now since the probability is the highest (0.8424410) among the three categories.

Conclusion and Discussion

In this paper, we introduced and studied a novel statistical approach to investigate the development of Chinese programs in U.S. higher institutions. The proportional odds ordinal logistic regression was conducted based on a sample of 201 American higher institutions which offer Chinese programs.

The estimated coefficients and intercepts from the model can be used to better understand the relationship between the Chinese program development and related factors like university enrollment, acceptance rate, etc. It also provides insights to higher institutions without Chinese program on the planning of potential new Chinese programs using predicted probabilities and classification. In addition, the model classified 70.15 % of the 201 higher institutions correctly, indicating that the model is appropriate for our study.

However, the model was developed in a relatively small data set and a smaller set of predictors. Larger samples and more predictors should be included in future study. In addition, typically, external validation of the proportional odds model is needed before it can be considered for application in practice. In our study, we tested the performance of our model by internal validation. Further methodological work is need for the continuation of future investigation.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: John Wiley and Sons.

Agresti, A. (2010). *The analysis of ordinal categorical data* (2nd ed.). New York: John Wiley and Sons.

Archer, K. J., & Williams, A. A. A. (2012). L1 penalized continuation ratio models for ordinal response prediction using high—dimensional datasets. *Statistics in Medicine, 31*(14), 1464-1474.

Armstrong, B. B., & Sloan, M. (1989). Ordinal regression models for epidemiological data. *American Journal of Epidemiology, 129*(1), 191-204.

Chang, K. E., Lan, Y. J., Chang, C. M., & Sung, Y. T. (2010). Mobile-device-supported strategy for Chinese reading comprehension. *Innovations in Education and Teaching International, 47*(1), 69-84.

Chen, H. Y., & Liu, K. Y. (2008). Web-based synchronized multimedia lecture system design for teaching/learning Chinese as a second language. *Computers & Education, 50*(3), 693-702.

Clogg, C. C., & Shihadeh, E. S. (1994*). Statistical models for ordinal variables.* Thousand Oaks, CA: Sage.

Dobuzinskis, A. (2011). *More U.S. students learning Chinese as school language programs expand*. Retrieved April 20, 2012, from http://www.huffingtonpost.com/2011/04/21/more-us-students-learning_n_852093.html

Goldberg, D., Looney, D., & Lusin, N. (2015). Enrollment in languages other than English in United States Institutions of higher education, Fall 2013. In *Modern Language Association*. Retrieved February 19, 2015, from http://www.mla.org/2009_enrollmentsurvey

Grant, S., & Huang, H. (2010). The integration of an online 3D virtual learning environment into formal classroom-based undergraduate Chinese language and culture curriculum. *Journal of Technology and Chinese Language Teaching, 1*(1), 2-13.

Graphiq. (2017). *Chinese language and literature.* Retrieved April 6, 2017, from http://colleges.startclass.com/d/o/Chinese-Language-and-Literature

Hilbe, J. M. (2009*). Logistic regression models*. Boca Raton, FL: Chapman& Hall/CRC.

Hosmer, D. W., & Lemeshow, S. (2013). *Applied logistic regression* (2nd ed.)*.* New York: John Wiley and Sons.

Huang, C. K., Lin, C. Y., & Chiang, Y. H. (2010). Incorporating competency-based blended learning in a Chinese language classroom: A Web 2.0 Drupal module design. *International Journal of E-Learning, 9*(4), 529-548.

Lin, M. C., & Wang, P. Y. (2008). A knowledge management system for Chinese language arts teachers. *British Journal of Educational Technology, 39*(5), 935-943.

Liu, X. (2009). Ordinal regression analysis: Fitting the proportional odds model using Stata, SAS and SPSS. *Journal of Modern Applied Statistical Methods, 8*(2), 632-645. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol8/iss2/30/

Liu, X. (2014). Fitting stereotype logistic regression models for ordinal response variables in educational research (Stata). *Journal of Modern Applied Statistical Methods, 13*(2), 528-545.

Nelley, F., Goldberg, D., & Lusin, N. (2010). Enrollments in languages other than English in United States institutions of higher learning. In *Modern Language Association*. Retrieved from http://www.mla.org/pdf/2009_enrollment_survery.pdf

Neely, B. (2011). *More Americans learn their ABCs in Chinese.* Retrieved April 20, 2012, from http://www.npr.org/2011/01/19/133031008/american-interest-in-learning-Chinese-skyrockets

Norris, J. (2016). Language program evaluation. *Modern Language Journal, 100*(Issue SI), 169-189.

O'Connell, A. A. (2006). *Logistic regression models for ordinal response variables*. Thousand Oaks: SAGE.

O'Connell, A. A., & Liu, X. (2011). Model diagnostics for proportional and partial proportional odds models. *Journal of Modern Applied Statistical Methods, 10*(1), 139-175. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol10/iss1/15/

OELA Newsline. (2005, September 23). *National Security Education Program (NSEP) Announces Major New Grant Award: Chinese K-16 Pipeline Project—University of Oregon and Portland Public Schools*. Retrieved May 15, 2007, from http://www.ncela.gwu.edu/newline/archives/2005/09/national_securi.html

Rogers, M. (2012). *Demand high for dual immersion programs in Utah: The Salt Lake Tribune*. Retrieved November 1, 2014, from http://www.sltrib.com/sltrib/news/.../language-programs-Chinese-dual.html.csp

US News and World Report. (2017). *Best colleges*. Retrieved April 6, 2017, from https://www.usnews.com/best-colleges

Welles, E. B. (2004). Foreign language enrollments in United States institutions of higher education, Fall 2002. *ADFL Bulletin, 35* (2-3), 7-26.