GRADE PERFORMANCE IN STATISTICS: A BAYESIAN FRAMEWORK

Lawrence V. Fulton, Texas State University, San Marcos, Texas, USA
Nathaniel D. Bastian, University of Maryland University College, Adelphi, Maryland, USA
R. Muzaffer Musal, Texas State University, San Marcos, Texas, USA

ABSTRACT

In this paper, we investigate questions involving the impact of a multitude of covariates and their interactions on the scores of a comprehensive exit exam from 121 undergraduate students in Texas State University via a hierarchical Bayesian mixture model. The model uses a mixture of Beta distributions, inflated for the purposes of modeling the behavior of students who are no-shows for the exam. We formulate the predictive probability distribution of letter grade test scores as well as the probability that a student will belong to a particular grade cluster given a set of covariate values.

Keywords: *Educational Research, Hierarchical Bayesian Modeling, Predictive Statistics, Performance*

1. INTRODUCTION

In this paper, we propose a hierarchical Bayesian mixture model to predict future performance and explain variations in the results of an introductory business statistics course exit assessment of Texas State University students. We conceptualize the *N* students' test scores as clustered in *K* performance groups where the average grades of the clusters are monotonically ordered indicating relative performance that go from the least to the highest performance group. We refer to these performance groups as grade clusters *i*. In addition to modeling the average grades in each cluster, we also model the probability that a particular student will belong to any of the *K* clusters. We note that each semester there exists a group of students who usually do not have a positive probability of passing the course and, therefore, do not show-up for the final assessment exam. The list of zeroes assigned to these students forms our first grade cluster. We propose a mixture model to incorporate the multi-peakedness of the distribution of grades implied by a *K* that is greater than 1. We use the Harmonic Mean measure of Newton & Raftery (1994) in determining *K*.

The hierarchical Bayesian modeling approach allows us to do several things we would not be able to do with a Generalized Linear Model. First, we can use information and make predictions on small sub-groups in our sample. Second, we can investigate a richer structure than is available using any frequentist version of the proposed models. For instance, we can obtain the probability distribution of the average grades in the grade clusters and provide a student with a probability distribution of his/her grade given various covariate values. We can also choose to treat some covariates as nuisance factors, integrate them out, and use only those that a model-user deems relevant while not over/under estimating the uncertainty about the coefficients of the chosen covariates. Furthermore, our model allows a realistic framework in which a student's probabilities of belonging to various clusters are negatively correlated, and we are able to introduce student level variation about the average grades in each cluster.

We provide a set of methods to help students and instructors to understand the relationship between various covariates and students' probabilistic grade distribution. From the perspective of policy-makers, it would be of interest whether a particular student group seems to perform better or worse than others for fixed values of the other covariates. For example, we can compare probabilistically a particular grade a Hispanic student will obtain compared to a student who is not classified as Hispanic. Our model can be used as a framework to draw a number of inferences as we model the average grade of each grade cluster as well as the probability that a student is going to belong to this cluster.

2. METHODS

In the first sub-section, we discuss the data source used in the analysis. We then describe the mixture model of test scores without any covariates. This allows an intuitive understanding of the general model. Next, we introduce the incorporation of information from the covariates in order to explain the uncertainty about the membership to a particular grade cluster, as well as the average grade in each grade cluster. In the final sub-section, we describe how we select K, the number of grade clusters.

2.1 Data Source

Our data consists of information that we obtained from 121 students. We used the Hawkes Learning Systems (HLS) to obtain the number of minutes each student logs in. This measure will serve as a proxy to the amount of time a student puts into class. Note that HLS logs a student out after 5 minutes of inactivity so that we would not have to worry about inactive students with high HLS minutes logged. We obtained age, grade point average (GPA) at the beginning of the semester, Hispanic status, and gender from the university registration system. This study was designated as exempt by the Institutional Review Board. We scale the final exam grade between 0 and 1 and illustrate it in Figure 1. This figure illustrates the multiple peaked nature of the grades. As mentioned above, the first peak of zeroes occur due to the students who are no-shows for the final exam. Note that we determine the number of peaks based on the measure of Harmonic Mean as explained below in the model section.

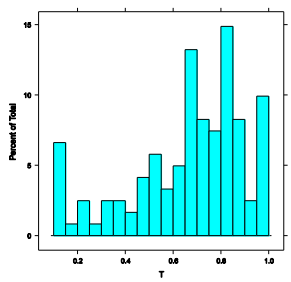

Figure 1: Histogram of 121 Student's Test Results

The covariates $Y_1 - Y_5$, represent Gender (1-female, 0-male), Age, Hispanic Status (1-Hispanic, 0-otherwise), Overall GPA, and Total Time Spent on the System, respectively. A student whose age (53) was over 8 standard deviations above the mean was removed from further analysis. There is substantial variation in all the covariates except the Hispanic status covariate. As is the current trend in business schools, the majority of the students are female. Also, the data indicates a rather wide dispersion among the students of GPA and total time spent in HLS minutes logged. There are a total of 71 female students and 29 Hispanics.

In Figure 2, we illustrate the relationships between grade distribution and the covariates. There are a total of 8 students who were no shows for the final exam. As the grades of the no-show students are fixed, we remove them for the purposes of illustrating the covariates' relationship to the final exam grades. For the continuous covariates, we add an Ordinary Linear Squares line to the plots. As can be seen from Figure 2, there seems to be a degree of linear trend of various strengths.

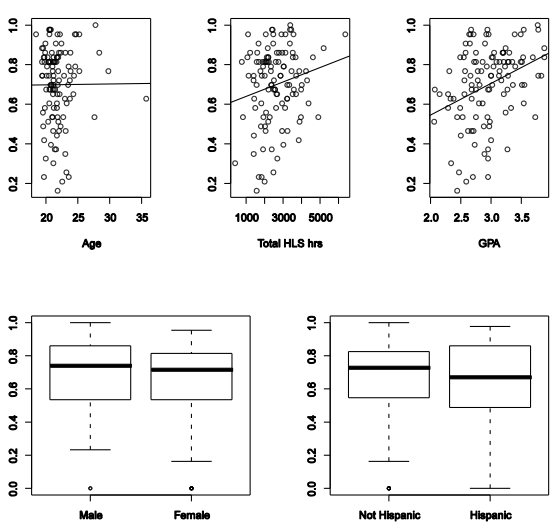

Figure 2: Test Results vs. Covariates

2.2 Model without Covariates

We use T to represent the final exam score of a student, scaled between 0 and 1, and assume that T is generated from one of K grade clusters where the first cluster is composed of the no-show students obtaining a score of 0. We chose to represent the likelihood function of the grade clusters with the beta distribution. We justify our choice due to the flexibility of the shape of the beta distribution with changes in its parameters and its support being bounded between 0 and 1. The Zero Inflated Mixture of Beta distribution can be represented as:

$$T = X_1 * 0 + \ldots + X_k * \beta((\kappa^{\mu} * \mu_k), \kappa^{\mu}(1-\mu_k)) + \ldots + X_K * \beta((\kappa^{\mu} * \mu_K), \kappa^{\mu} * (1-\mu_K)) \quad (1)$$

In Equation 1, the **X** vector of length K has one of its elements equal to 1 and the rest 0. It is straightforward to model this vector as a multinomial distribution where the total number of trials is equal to one.

$$\mathbf{X} \sim Multinomial(\phi, 1) \quad (2)$$

In Equation 2, the elements of K length vector, ϕ, represent the distribution of probabilities that a student is going to belong to one of K grade clusters. We choose the Dirichlet distribution to represent this uncertainty about ϕ parameters. We parameterize this Dirichlet distribution such that the K elements sum up to one implying that α_k in Equation 3 is the expected probability that a student will belong to k'th grade cluster. Due to our parameterization, note that K elements of the Dirichlet distribution are negatively correlated, which is necessary to represent the probabilities of a discrete probability distribution. When coupled with this distribution's flexibility, the Dirichlet distribution is a natural choice in modeling vector ϕ.

$$\phi \sim Dirichlet(\kappa^{\alpha} * \alpha_1, \ldots, \kappa^{\alpha} * \alpha_K) \quad (3)$$

where $\sum_{k=1}^{k=K} \alpha_k = 1$ and κ^{α} is the precision term of the Dirichlet distribution.

Due to the definition of grade clusters, we should have an order of the random variables representing their distribution. We use the ordered Dirichlet distribution to represent the uncertainty about the average grades in each cluster. The ordered Dirichlet is a natural choice as it allows us to incorporate the structure that there is a monotonic order to the expected values of the average test scores in the grade clusters, the μ terms in Equation 4. Note that the index of μ in Equation 4 starts from 2 as the index 1 refers to the least performing students whose grades are fixed and do not need to be modeled. Details of the Dirichlet and ordered Dirichlet distribution and its various parameterizations are discussed in Mazzuchi et al. (1997) and Mazzuchi et al. (1998).

$$\mu \sim Ord.Dirichlet(\kappa^{\gamma} * \gamma_2, \ldots, \kappa^{\gamma} * \gamma_K) \quad (4)$$

where $\sum_{i=3}^{i=K} \gamma_k - \gamma_{k-1} = 1$ and κ^{γ} is the precision term. Note that $\gamma_k - \gamma_{k-1}$ is the expected grade change between grade cluster k and $k - 1$. The k index in Equation 4 starts from 3 since the very first grade cluster group has its value fixed as it represents the no-show students and therefore is not part of a probability distribution. Also, γ_2 represents the expected grade change of the students whose grades are clustered to represent the next least performing students after the no-show grade cluster.

Ordered Dirichlet distribution is similar to the Dirichlet but the random variables in the distribution are monotonically ordered. This allows us to introduce identifiability to the mixture distributions. We summarize this model in Equation 5 below:

$$\begin{gathered} T = X_1 * 0 + \ldots + X_k * \beta((\kappa^{\mu} * \mu_k), \kappa^{\mu}(1-\mu_k)) + \ldots + \\ X_K * \beta((\kappa^{\mu} * \mu_K), \kappa^{\mu} * (1-\mu_K)) \\ \mathbf{X} \sim Multinomial(\phi, 1) \\ \mu \sim Ord.Dirichlet(\kappa^{\gamma} * \gamma_2, \ldots, \kappa^{\gamma} * \gamma_K) \\ \phi \sim Dirichlet(\kappa^{\alpha} * \alpha_1, \ldots, \kappa^{\alpha} * \alpha_K) \\ \gamma_k - \gamma_{k-1} \sim Dirichlet(1, \ldots, 1); \alpha \sim Dirichlet(1, 1 \ldots, 1) \\ \kappa^{\mu} \sim Gamma(1, .1); \kappa^{\gamma} \sim Gamma(1, .1); \kappa^{\alpha} \sim Gamma(1, .1) \end{gathered} \quad (5)$$

2.3 Model with Covariates

The model we develop in this subsection allows us to make inferences on the effects of various covariates on a students' probability of belonging from the highest to the lowest performance grade clusters. We use covariates in an N by M matrix $\mathbf{Y}$ to model the expected probabilities, where N is the number of students and M is the number of covariates. Namely, we model the elements of vector ϕ_k in Equation 5. We set the expected probability that student n is going to belong to grade cluster k, $\alpha_{n,k}$, equal to the ratio of a function of the M covariates for that grade cluster to the sum of the functions associated with all the grade clusters. This parameterization ensures that $\alpha_{n,k}$ has a support between 0 and 1 and $\sum_{k=1}^{k=K} \alpha_{n,k} = 1$. We formulate this in Equation 6. The value of the function for a particular grade cluster k for student n, is represented as $\eta_{n,k}$. The $\eta_{n,k}$ terms can be greater than or equal to 0 which is ensured by the formulation shown in Equation 7. The log of $\eta_{n,k}$ has a support between positive and negative infinity and is the sum of the common intercept term in grade cluster k, χ^{η}_{k}, and the multiplication of the M coefficients with their respective covariates. In order to make our parameter estimation of the intercepts, χ^{η}, and coefficient of the covariates, ρ^{η} in $\eta_{n,k}$, more efficient, we set the parameters associated with the K'th grade cluster equal to 0 and use Equation 6 to make the other parameters relative to them. This formulation allows us to estimate one set of parameters more than we otherwise would by including the coefficients associated with the K'th cluster.

$$\left(\alpha_{n,k}\right) = \left(\frac{\eta_{n,k}}{1+\sum_{k=1}^{K-1} \eta_{n,k}}\right) \tag{6}$$

In Equation 6, we fix the coefficients of (η_K) equal to zero. This implies that the remaining coefficients are estimated relative to this last set of coefficients. The parameter $\eta_{n,k}$ is computed with a linear equation and a common intercept term.

$$log\ (\eta_{n,k}) = \chi^{\eta}_{k} + \rho^{\eta}_{k} * Y_{n,1} + \ldots + \rho^{\eta}_{k} * Y_{n,M} \tag{7}$$

Our choice of priors as normal distributions with mean 0 and variance 10 might seem less than diffuse. Note, however, that the random variable from this normal distribution is exponentiated and is a component in a ratio that is relative to the K'th category. Note that in Equation 3, the K length vector ϕ_n and μ_n represent the probability and the average grade that student n will belong to a particular grade cluster and the average grade in that cluster, respectively. κ^{α} represents the common precision term. We use the covariate information in $\mathbf{Y}$ to model the ϕ and μ vector using the well-known log link function. We extend the model in order to model the μ terms. Note that $\mu_{n,k}$ represents the expected grade student n receives in grade cluster k. We again use the ordered Dirichlet distribution.

$$\mu_{n,k} \sim Ord.Dirichlet(\kappa^{\gamma} * \gamma_{n,2},\ldots,\kappa^{\gamma} * \gamma_{n,K}) \tag{8}$$

where the γ parameters of the ordered Dirichlet distribution are monotonically increasing from $\gamma_{n,2}$ to $\gamma_{n,K}$, the κ^{γ} term is again the common precision term. We use the familiar link function:

$$\left(\gamma_{n,k}\right) = \left(\frac{\psi_{n,k}}{1+\sum_{k=2}^{K-1} \psi_{n,k}}\right) \tag{9}$$

The set of $\psi_{n,k}$ parameters are computed again with a linear equation,

$$log\ (\psi_{n,k}) = \chi^{\psi}_{k} + \rho^{\psi}_{k,1} * Y_{n,1} + \ldots + \rho^{\psi}_{k,M} * Y_{n,M} \tag{10}$$

where a student obtains a zero test score with probability α_1 and belongs to test score group n with probability α_k. We use the information in the covariates ($\mathbf{Y}$) in modeling the probability that a student is going to belong to a particular grade group as well as the parameters of the grade distributions. This is done through the well-known log link function. We fix the values of the covariates associated with α_K and μ_K equal to 0. This implies that the values of the remaining parameters are computed relative to k'th group. We use diffuse priors to let the data drive the learning process on the parameters. The number of grade clusters K is determined by making use of Harmonic Mean, where $\sum_{k=1}^{K} \mu_k = 1$ and $\sum_{k=1}^{K} \alpha_k = 1$:

$$
\begin{gathered}
T_n = X_{n,1} * 0 + \ldots + X_k * \beta((\kappa^{\mu} * \mu_k), \kappa^{\mu}(1-\mu_k)) + \ldots + \\
X_K * \beta((\kappa^{\mu} * \mu_K), \kappa^{\mu} * (1-\mu_K)) \\
\mathbf{X} \sim Multinomial(\phi, 1) \\
\mu \sim Ord.Dirichlet(\kappa^{\gamma} * \gamma_2, \ldots, \kappa^{\gamma} * \gamma_K) \\
\phi \sim Dirichlet(\kappa^{\alpha} * \alpha_1, \ldots, \kappa^{\alpha} * \alpha_K) \\
(\alpha_{n,k}) = \left(\frac{\eta_{n,k}}{1+\sum_{k=1}^{K-1}\eta_{n,k}}\right); (\gamma_{n,k}) = \left(\frac{\psi_{n,k}}{1+\sum_{k=2}^{K-1}\psi_{n,k}}\right) \\
log\,(\eta_{n,k}) = \chi^{\eta}_{k} + \rho^{\eta}_{k,1} * Y_{n,1} + \ldots + \rho^{\eta}_{k,M} * Y_{n,M} \\
log\,(\psi_{n,k}) = \chi^{\psi}_{k} + \rho^{\psi}_{k,1} * Y_{n,1} + \ldots + \rho^{\psi}_{k,M} * Y_{n,M} \\
\chi^{\eta}_{1,\ldots,K-1} \sim Normal(0, 0.1);\ \rho^{\eta}_{1,\ldots M,\ldots,K-1} \sim Normal(0, 0.1); \\
\chi^{\psi}_{2,\ldots,K-1} \sim Normal(0, 0.1);\ \rho^{\psi}_{2,\ldots M,\ldots,K-1} \sim Normal(0, 0.1); \\
\chi^{\eta}_{K} = 0;\ \rho^{\eta}_{K,} = 0;\ \chi^{\psi}_{K} = 0;\ \rho^{\psi}_{K} = 0 \\
\kappa^{\mu} \sim Gamma(1, .1);\ \kappa^{\gamma} \sim Gamma(1, .1);\ \kappa^{\alpha} \sim Gamma(1, .1) \\
\text{where } \sum_{k=1}^{K}\mu_k = 1 \text{ and } \sum_{k=1}^{K}\alpha_k = 1
\end{gathered}
\tag{11}
$$

2.4 Selecting *K*

Note that we can choose any integer that is greater than or equal to one for *K*. With each possible decision, we would be defining a new model. A measure that is used to select between different models is the Harmonic Mean Index. The Harmonic Mean Index computes the prior predictive of the data, $p(T)$. This measure allows us to compare different models where the higher values of $p(T)$ imply a better fit to the data. In computing $p(T)$ we utilize the property below:

$$p(T) = \left[\frac{1}{S}\sum_{s=1}^{s=S}\sum_{i=1}^{i=K}\frac{1}{p(T_i|\theta)}\right]^{-1} \tag{12}$$

3. RESULTS AND DISCUSSION

We applied our models using OpenBugs (Spiegelhalter et al., 2009), and we used the R package in analyzing convergence, constructing the posterior predictives, and conducting remaining analysis (R Development Core Team, 2009; Plummer et al., 2010).

As suggested by Gelman & Hill (2007), in order to circumvent numerical difficulties, we constrain for each i, the probability that a student belongs to a certain grade cluster to be between 0.001 and 0.99. We also constrain the grade difference between the grade clusters to the same values. Furthermore, we truncate the average value of the second grade cluster to be at least 0.1 based on historical observations.

In both of the models, we monitor Gelman and Rubin's $\hat{R}$ statistic (Rubin & Gelman, 1992) as well as Brooks and Gelman multivariate $\hat{R}$ convergence statistics (Gelman & Brooks, 1997) for each of the parameters in the three independent chains. In the models that do not make use of covariates, convergence is reached quickly and all of the parameters have $\hat{R}$ values between 1 and 1.01 as point estimates. As could be expected, the model with an additional layer of hierarchy and with the covariates took longer to reach convergence. We ran the MCMC simulation until all of the parameters have a point estimate and the multivariate $\hat{R}$ value that is less than 1.2.

We investigate six models that have no covariates in order to determine the optimum number of grade clusters, *K*. We use the Harmonic Mean Index introduced in Equation 12 as our measure of fit. It can be seen in Table 1 that the model with 5 grade clusters seems to provide the best trade-off between model complexity and fit.

Table 1: Number of Clusters vs. Harmonic Mean

Number of Grade Clusters	Harmonic Mean
K = 2	5.14
K = 3	5.68
K = 4	5.97
K = 5	6.07
K = 6	5.90

We apply the model we introduced in Equation 11, where we model the expected grade in each grade cluster for each student in addition to the set of probabilities that a student belongs to a particular grade cluster. Recall that the average grade of student *n*, in grade cluster *k*, $\gamma_{n,k}$, is computed as the ratio

of $\psi_{n,k}$ divided by the sum of all the K ψ terms. Furthermore, recall also that we model the logarithm of $K-1$ ψ and η terms with a linear equation involving a χ_k term and $M * \rho$ term, where all of the K'th grade cluster's coefficients are set to 0 to improve convergence and efficiency. This means that all the χ and ρ coefficients are estimated relative to the K'th grade cluster values. As mentioned above, we have both M and K as equal to 5.

In Figure 3, we show the distribution of letter grades for 12 sets of hypothetical, non-Hispanic, male students with an average GPA. We divide the HLS minutes studied into 12 equidistant points within two standard deviations of the average HLS minutes logged. As can be expected, we observe in Figure 3 the letter grade distributions of the hypothetical students slowly get a more right skewed distribution as the number of HLS minutes logged increases.

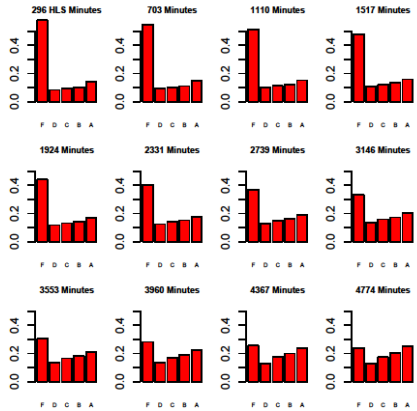

Figure 3: Posterior Predictive Letter Grade Probability Distributions of 12 Sets of Students

Using these models, we can compare multiple students with various characteristics. Assume four hypothetical students: Student 1 has average age, a GPA that is 2 standard deviations below the mean (2.05), utilizes the HLS 3 standard deviations above the average minutes (5893.2) and is a Hispanic female; Student 2 also has the average age, a GPA that is 2 standard deviations above the average (3.81), utilizes the HLS an average number of minutes (2535) and is non-Hispanic male; Student 3 has average age, average GPA, utilizes the HLS 3 standard deviations above the average minutes and is a Hispanic female; Student 4 has average age and a GPA that is 1 standard deviation above the mean (3.37). In Table 2, we obtain the posterior predictive probability distributions of these students.

Table 2: Posterior Predictive Letter Grade Probability Distributions of Four Students

	Student 1	Student 2	Student 3	Student 4
F	0.562	0.346	0.361	0.269
D	0.070	0.109	0.092	0.092
C	0.081	0.126	0.114	0.122
B	0.101	0.155	0.144	0.162
A	0.186	0.264	0.288	0.354

4. CONCLUSIONS

In this paper, we described a Bayesian methodology to predict the final exam scores of a set of students. Our methods are based on the assumption that every semester there are a fixed number of grade clusters that each student belongs to. We demonstrated how to use the Harmonic Mean Index to identify the number of grade clusters that provides the "best" tradeoff between complexity and fit. We constructed a realistic model to predict the grade distribution of students. We used numerous covariates to predict the probability that a student will belong to a particular grade cluster as well as the score he/she will get in that cluster. Our findings indicate that even though there is still a large posterior variance, GPA and HLS logged minutes are especially important. This is consistent with our belief that a high GPA or spending time on HLS does not guarantee learning actually occurs. These methods can be used to explain to the students that even though HLS minutes logged definitely helps, it is not enough by itself.

5. REFERENCES

Gelman, A. and Brooks, S.P. (1997) General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7, 434-455.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2003). *Bayesian Data Analysis*. Chapman and Hall CRC.

Gelman, A. and Hill, J. (2007). *Data analysis using regression and multi-level/hierarchical models.* Cambridge University Press.

Gelman, A., Meng, X.L. and Stern, H. (1996) Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, pages 733-807.

Mazzuchi, T.A., Soyer, R. and Erkanli, A. (1998) Bayesian computations for a class of reliability growth models. *Technometrics*, 40, 14-23.

Mazzuchi, T.A., Soyer, R. and van Dorp, J.R. (1997) Sequential inference and decision making for single mission systems development. *Journal of Statistical Planning and Inference*, 62, 207-218.

Newton, M.A. and Raftery, A.D. (1994) Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, 56(1), pp. 3-48.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2010) coda: *Output analysis and diagnostics for MCMC.* R package version 0.13-5.

R Development Core Team. (2009) *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rubin, D.B. and Gelman, A. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-511.

Spiegelhalter, D., Thomas, A., Best, N. and Lunn, D. (2009) The bugs project: Evolution, critique, and future directions. *Statistics in Medicine*, 28, 3049-3067.

Author Profiles:

Dr. Lawrence V. Fulton holds a Ph.D. degree from the University of Texas at Austin. He currently teaches quantitative methods in the McCoy School of Business at Texas State University.

Nathaniel D. Bastian holds a M.Sc. degree from Maastricht University in the Netherlands. He currently teaches mathematics and statistics for the University of Maryland University College.

Dr. R. Muzaffer Musal holds a Ph.D. degree from George Washington University. He currently teaches quantitative methods in the McCoy School of Business at Texas State University.