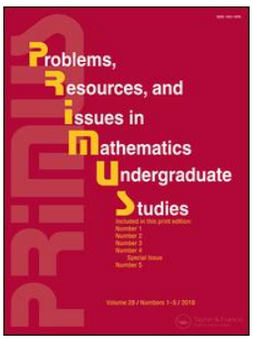

PRIMUS

Problems, Resources, and Issues in Mathematics Undergraduate Studies

ISSN: 1051-1970 (Print) 1935-4053 (Online) Journal homepage: https://www.tandfonline.com/loi/upri20

The Impact of Academically Homogeneous Classrooms for Undergraduate Statistics

Dusty Turner, James Pleuss & Christopher Collins

To cite this article: Dusty Turner, James Pleuss & Christopher Collins (2020): The Impact of Academically Homogeneous Classrooms for Undergraduate Statistics, PRIMUS, DOI: 10.1080/10511970.2019.1710629

To link to this article: https://doi.org/10.1080/10511970.2019.1710629

Accepted author version posted online: 04 Jan 2020.
Published online: 30 Jan 2020.

PRIMUS, 0(0): 1–13, 2020
ISSN: 1051-1970 print / 1935-4053 online
DOI: 10.1080/10511970.2019.1710629

The Impact of Academically Homogeneous Classrooms for Undergraduate Statistics

Dusty Turner, James Pleuss, and Christopher Collins

Abstract: In the continual pursuit of classroom learning effectiveness, researchers and educators aim to develop strategies that improve student performance and learning. One such strategy is to create academically homogeneous environments where students are grouped into classes based on their preconceived academic ability. The research team tests this common assertion through an experiment at The United States Military Academy at West Point. Students are placed in either ability- or randomly grouped sections of the academy's mandatory introduction to probability and statistics course, where their ability projections are based on mathematical modeling of a student's historical academic performance. A quantitative analysis of student performances across major graded events indicates that there is not a significant difference in student performance between ability- and randomly grouped sections. This is the first robust experiment of this nature at the undergraduate level and provides useful insight to educators and administrators alike.

Keywords: Statistics, undergraduate, education, pedagogy

1. INTRODUCTION

Educators always work to develop strategies to help their students become better learners. One commonly used and often divisive tactic is the implementation of advanced classrooms where students of higher ability work together in what this research calls an academically homogeneous

Address correspondence to Dusty Turner, Department of Mathematical Sciences, The United States Military Academy, 646 Swift Road, West Point, NY 10996, USA, Center for Army Analysis, 6001 Goethals Road, Fort Belvoir, VA22060, USA. E-mail: dusty.s.turner.mil@mail.mil.

environment. William Viar [7] and Dr. Robert Slavin [6] explored "tracking" of students by placing them in tracks for the streamlined curriculum that met the student's identified skill set as early as elementary school. They noted some of the characteristics of these strategies and the types of ability grouping used.

Dr. Tom Loveless, the most referenced author in cases against ability grouping and tracking [4], notes that a large portion (about 50%) of surveyed educators wanted more ability grouping in K-12 classrooms. Additionally, only 40% of those surveyed said they believed that heterogeneously mixed classrooms would improve a student's education. As those surveyed were further filtered to only include middle school educators, those that preferred ability grouping rose to 66%, indicating that they saw the benefit of pairing students with similar abilities. Loveless concludes, however, that tracking, and to a lesser extent ability grouping, "fosters race and class segregation", potentially harming students' self-esteem, and leading to a self-fulfilling prophecy for most students. In a complementary study, Dr. William Carbonaro [1] supported Dr. Loveless's findings as it was found that students' effort is strongly correlated to the track they are placed in, with the lower tracks exhibiting lower academic effort. It is worth noting again that these studies were all conducted on students in elementary and middle school education and yielded anecdotal results.

There are very few studies that look at how ability sectioning impacts undergraduate learning. This is largely because most undergraduate institutions have large class sizes and give students the freedom to choose their time and instructor. One study that overcame these challenges was conducted by Army Lieutenant Colonel (LTC) Randall Hickman [2] who compared the results of Multivariate Calculus exams between years where students were grouped by ability one semester and were assigned at random the next. His results were mixed, and since the experiment was conducted over two different semesters with different instructors and slightly different exams, it lacked a design that was conducive to drawing substantiated conclusions.

In this paper, we build off of LTC Hickman's work to answer the question: "Do introductory statistics students learn better when they are in classrooms with others of similar ability?" Receiving support from the Department of Mathematical Sciences leadership, the authors designed and implemented a study to answer this question as indicated in Table 1.

Students at The United States Military Academy take a mandatory sequence of three core math courses as indicated in Table 1, the last of which is MA206: Introduction to Probability and Statistics. MA206 is a calculus-based probability and statistics course with a co-requisite of single variable calculus for its students. Academies are uniquely suited to rigorously conduct such a research experiment for the following reasons:

Table 1. Math course tracks taken for cadets during AY 17-01 and AY 17-02. The Standard STEM (Science, Technology, Engineering, and Mathematics) and Standard Non-STEM tracks are the most common tracks for students

Track	Courses	Cadets
Standard Non-STEM	MA103, MA104	454
Standard STEM	MA103, MA104, MA205	320
STEM w/ MA103 Validation	MA104	23
Remedial to STEM w/ Validation	MA100, MA104, MA205	10
Advanced Non-STEM	MA153	8
Advanced STEM	MA153, MA255	197
Advanced to Standard	MA153, MA205	7

1. Course leadership can exercise control over the students in terms of what hours they take a particular course and what section they are assigned to.
2. Most instructors teach multiple sections of the same course, allowing for more effectual control/treatment groups.
3. Lessons are established by course leadership and followed by each instructor for identical teaching content.
4. Major graded events are identical for all students and are centrally graded (i.e., all instructors come together to grade, each with their own portion of the exam).

Proponents of the typical heterogeneous (randomly selected) classroom may cite a benefit to diversity in the classroom or an increase in student-to-student teaching. We desire to know if students will perform better if they are grouped with other students of similar ability. The benefits could be that teachers will be able to better prepare for students of the same ability, weaker students will be more likely to ask questions if they do not feel intimidated by stronger students, and stronger sections will be able to go deeper into the material because they will not be held back by weaker students. In the following pages, we examine these assumptions.

2. EXPERIMENTAL DESIGN

2.1. Ability Group Determination

To conduct this experiment, a metric defining student ability must be used. This research uses projected performance in the course as determined by statistical models built with the past course data to define ability. The predictive models built used data from 1023 students from the fall semester of 2016 and spring semester of 2017 (Academic Year (AY) 17-01 and AY 17-02, respectively). We separated our data into a testing and training set consisting of 75% and 25% of the available observations, respectively. From there we created

Table 2. Independent variables used for each cadet and the description of these variables

Independent variable	Description
Modeling Grade	Final Grade of MA103 or MA153
Single-Variable Calculus Grade	Final Grade of MA104 or MA205
Multi-Variable Calculus Grade	Final Grade of MA153 or MA255
Core Math Average	Average of Core Math GPA
CQPA	Cumulative Academic Grade Point Average
CCPS	Cumulative Overall Grade (Academic/Military/Physical)
Rock	Whether or not a cadet took remedial "rock" math

three different types of models: a Random Forest, a LASSO, and a Linear Regression Model, each developed through 10-Fold Cross-Validation from the CARET (Classification And REgression Training) package in R [3]. We used a model ensemble (combining them all in a linear combination) to limit the errors made by any particular model. We use a cut off of $\alpha = 0.05$ for all hypothesis tests.

Inputs into our model were the track of the previous math courses taken (Table 1) and academic performance scores for each cadet (Table 2). These variables represent the factors for possible inclusion in the predictive model.

Based on their previous math experience, West Point offers certain students an opportunity to test out of a course by achieving 80% or better on the course final exam. This validation caused missing values for some of the Single-Variable Calculus and Modeling Grades (2.26% and 0.1% of the observations, respectively). We imputed these scores using a K-Nearest-Neighbor clustering method with "k" $= 5$.

Table 3 provides a summary of the fits of the models used to project MA206 course average. Instead of selecting one model, we chose an ensemble method so that no single model, if it made systematically biased estimates, would dominate the predictions. While there are many complicated ways to ensemble multiple models, we averaged the response of the three models versus other more complex techniques. This helped maintain interpretability while giving equal weight to each model [5].

2.2. Control and Treatment Groups

In the semester pertinent to this study, AY 18-01, MA206 consisted of 580 cadets spread across four course hours. We split the population into a control (random) group, consisting of students assigned to classrooms at random, and a treatment (ability) group, consisting of students assigned to classrooms by their ability. Table 4 describes the breakdown of how cadets were assigned to each hour. The two groups (ability and random) had a similar overall academic performance, as shown by the summary statistics in Table 5. This reveals that

Table 3. Summary of the fits of the course average projection models. RMSE is the residual mean squared error and gives the sum of squared errors for each of the observations. Lower numbers indicate less overall error from actual course grades to the predicted ones. R^2 is the coefficient of determination and reveals the proportion of variation in course score accounted for by the model. The closer the number to one, the more variation the model takes into account

Model	RMSE	R^2
Linear Regression	0.2631	0.9308
Regularization	0.2626	0.9310
Random Forest	0.3148	0.9010
Ensemble Model	0.2663	0.9291

Table 4. Experimental design technique for cadet assignment while accounting for class day/time. The United States Military Academy has a Day 1/Day 2 schedule instead of the more traditional M-W-F, T-Th schedule. B and H hours (as well as C and I hours) are offered at the same time on alternating days

Hour	Day	Time	Random/Ability	Cadets
B	1	840–935	Ability	145
C	1	950–1045	Random	135
H	2	840–935	Random	157
I	2	950–1045	Ability	143

Table 5. Summary statistics of the students in ability versus random hours. GPA is the student's GPA prior to taking MA206 represented with the mean plus or minus one standard deviation; multivariable calculus is the proportion of students in the group that has taken a multivariable calculus course (all students take single-variable calculus prior to MA206); predicted grade is the grade that our model predicted that student would receive in MA206, represented with the mean plus or minus one standard deviation

Group	GPA	MultiVariable Calculus	Predicted Grade
Ability	3.01 ± 0.573	0.309	0.828 ± 0.078
Random	2.98 ± 0.517	0.367	0.822 ± 0.069

the groups were structured in a way that limits bias associated with how they might perform in MA206.

2.3. Teaching Constraints

Instructor teaching schedules and administrative duties prevented every instructor from teaching both a random (control) section and a randomly assigned ability (treatment) section. Additionally, to reduce the impact of

Table 6. Instructor allocation of sections. B and H hours are offered from 840 to 935 on alternating days. C and I hours are offered from 950 to 1045 on alternating days

Instructor ID	B hour	I hour	C hour	H hour
1	Ability 5	Ability 5	Random	Random
2		Ability 6		Random
3	Ability 8	Ability 8	Random	Random
4	Ability 1		Random	Random
5		Ability 1		
6	Ability 6			
7				Random
8	Ability 3	Ability 3	Random	Random
9	Ability 7	Ability 7	Random	Random
10	Ability 2	Ability 2	Random	Random
11	Ability 4	Ability 4	Random	Random
12			Random	
13	Ability 9	Ability 9	Random	Random

preparing for multiple classes of differing student abilities, instructors that teach two ability sections were assigned classrooms with the same level of students. While these constraints add the possibility of confounding in our results, we control for this by including ability level in our model.

The instructor breakdown is shown in Table 6, where each instructor has a unique ID number and the projected ability of a section decreases as the corresponding number increases. West Point's vision for instruction includes small class sizes with more instructor involvement. This results in a class section typically containing 16–18 students. The nine ability sections for B and I hours come from the number of sections required to teach all the students assigned to that hour by the registrar. Each student taking MA206 in B or I hours was assigned to the appropriate ability group (section) based on their predicted performance using the model in Section 2.1. For example, of students taking MA206 during B hour, the 16–18 with the highest predicted score made up a single section we call Ability 1. Conversely, Ability 9 for B hour is the section containing the 16–18 students taking MA206 during B hour with the worst projected scores.

To further limit the impact of the instructor as a confounding variable, we solely analyze overall grade average on course-wide graded events. This includes three mid-term tests and one comprehensive final exam.

As a final note about experimental design relating to instructors, we did not execute this as a blinded study. Teachers were aware of the make-up of the students in their classrooms. It would not have been possible for teachers to remain unaware of their students' talents as teachers have access to students' previous performance. Furthermore, students in ability sections, though not formally told, were likely aware that they were in sections of similar projected ability.

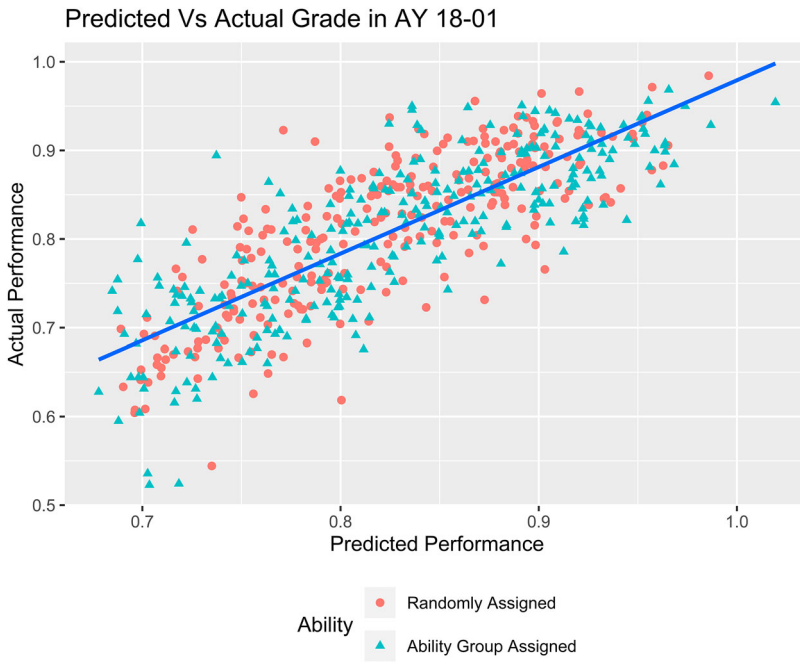

Figure 1. Predicted vs. actual scores by grouping.

3. RESULTS

3.1. Model Results

At the conclusion of the experiment, we analyzed course performance in several different ways to determine if student learning improves when placed in classrooms with other students of similar ability. We conducted multiple two-sample tests between students of the same ability level, split across whether the students were in the ability or random groups. Second, we created a linear model to predict their end-of-course average using their experimental group and controlling for other confounding variables.

In summary, this section shows how our model explains 61% of the variation in MA206 scores (R^2). In Figure 1 you can see the general trend in predicted vs. actual scores.

3.2. Multiple Pairwise Comparisons

The first of our two pairwise comparisons tests determine if the mean course average of those sectioned by ability is different than those randomly assigned. Summary statistics of these groups are in Table 7.

Table 7. Descriptive statistics for both groups

	Random	Ability
Cadets assigned	272	283
Mean grade	0.8104	0.8060
Standard deviation	0.0876	0.0906

Table 8. Descriptive statistics when grouped by ability.

Ability level	Random	Ability	*P*-value	Significant
Ability 1	0.903	0.901	0.883	No
Ability 2	0.878	0.872	0.552	No
Ability 3	0.869	0.881	0.299	No
Ability 4	0.847	0.842	0.686	No
Ability 5	0.836	0.844	0.573	No
Ability 6	0.789	0.771	0.225	No
Ability 7	0.762	0.753	0.563	No
Ability 8	0.735	0.706	0.049	Yes
Ability 9	0.676	0.688	0.483	No

The 95% confidence interval of the true mean difference in students randomly assigned versus those sectioned by ability is $(-.0104, 0.019307)$. Because of this, we cannot conclude that there is a difference between the mean of these two groups.

We also considered pairwise comparisons across the nine ability levels. Table 8 shows the means in performance between the students in each ability group and those assigned randomly that would have been in the same ability group. The last column denotes whether the difference between the groups is statistically significant or not.

It appears there is a significant difference at Ability Level 8 between the randomized and the ability-grouped sections. However, after applying even the most liberal corrections for multiple hypothesis testing, the finding is irrelevant. It is likely that this finding is due to expected randomness in the data as indicated by the rest of the ability groups rather than by a systematic difference in performance for students in Ability Level 8.

To further support the evidence that there is no measurable difference in student learning when students are sectioned by ability, Figure 2 shows the performance level of each group side-by-side based on their inherent student ability level. There is no discernible difference between the two types of classrooms and certainly no evidence that the ability sections out-perform their random counterparts. In fact, most of the ability sections performed worse on average than the randomly assigned students.

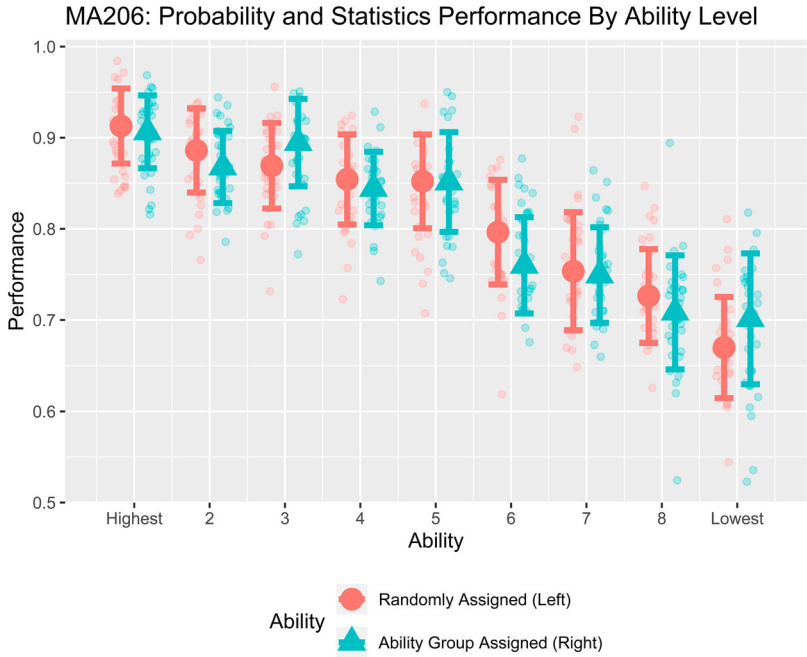

Figure 2. The difference in performance by ability and group. The middle points of the error bars are the means with the error bars representing two standard deviations from the respective means.

Table 9. Summary of the linear model using predicted grade and whether a student was in ability or random section. The coefficient for "Random" indicates the expected boost in student performance if a student was in a randomly assigned section. The coefficient for "Ability" is explained similarly

Covariate	Coefficient	*P*-value	Significant
Predicted Grade	0.980	< 0.001	Yes
Random	0.004	0.869	No
Ability	-0.006	0.822	No

3.3. Testing Effect of Sectioning by Ability

To determine the significance of sectioning by ability, we built two linear models; the first predicts student performance by the predicted score (to account for a student's ability) and whether the student was in a random or ability section. The output from this linear model is shown in Table 9. The final column states that while the predicted grade is significant (which we would expect), whether the student was in ability or random section did not impact their performance significantly.

Table 10. Summary of the linear model with ability group and random/ability included as factors.

Covariate	Coefficient	*P*-Value	Significant
Predicted Grade	0.9376	< 0.001	Yes
Ability 1 Random	0.0289	0.861	No
Ability 2 Random	0.0403	0.799	No
Ability 3 Random	0.0494	0.749	No
Ability 4 Random	0.0530	0.724	No
Ability 5 Random	0.0638	0.662	No
Ability 6 Random	0.0421	0.766	No
Ability 7 Random	0.0376	0.784	No
Ability 8 Random	0.0332	0.802	No
Ability 9 Random	0.0082	0.949	No
Ability 1 Ability	0.0122	0.942	No
Ability 2 Ability	0.0172	0.915	No
Ability 3 Ability	0.0515	0.743	No
Ability 4 Ability	0.0371	0.807	No
Ability 5 Ability	0.0671	0.648	No
Ability 6 Ability	0.0185	0.896	No
Ability 7 Ability	0.0274	0.842	No
Ability 8 Ability	0.0132	0.919	No
Ability 9 Ability	0.0270	0.829	No

The second linear model answers the question of whether a student may benefit from ability sectioning. We did this by including a covariate term for every ability level and randomize/non-randomized section combination. From these results in Table 10, we again see that none of the factors that represent a student's ability with regard to our experiment carries significance in explaining a student's performance. The only predictor with any significance is that which represents the student's projected score. This corroborates all other findings in our experiment.

4. DISCUSSION AND CONCLUSION

4.1. Limitations of the Experiment

While this experiment provides a wealth of quantitative analysis and observations, the breadth of its application is limited. The experiment only looks at an introduction to statistics course. Further research is necessary to determine the relevance to other undergraduate fields of study. The implications of the statistical results are limited to this type of course. We would anticipate, however, other STEM courses would yield similar results. A humanities course, for instance, may conduct a similar study with different results because on the

nature of its course structure. Another limitation is in the constraints discussed in Section 2.3. We were not able to completely randomize instructor assignment for the ability groups, introducing some hierarchical clustering effects. We were able to control for this effect by including instructor in our models, which takes this into consideration in our results.

4.2. Strengths of the Experiment

The strengths of this study far outweigh its limitations by offering a quantitative look into the impact of academically homogeneous classrooms in an undergraduate setting. To the authors' knowledge, this study is the first one to offer a scientific and experimentally robust study on the undergraduate population. The authors believe that similar statistics courses executed under the same conditions would yield similar results ; however, more experiments are necessary to validate this conjecture. The ensemble model used to predict student performance is much stronger than simply using a student's GPA as its ability measure, giving more credence to the results of the experiment. A final strength of the study is the breadth of approaches used to evaluate the results. Looking at the populations (ability vs. random) at large is useful, but doing a pairwise comparison of each ability level provides the reader more granular insight to how each level of student ability is impacted by the study.

4.3. Concluding Remarks

The goal of this research is to assess the consequence of grouping undergraduate statistics students by their ability. The results provide informed feedback to the Department of Mathematical Sciences leadership on whether this technique should be expanded or perhaps discontinued. The research team hypothesized a positive impact of sectioning students by their mathematical ability. However, based on the quantitative findings from this research, the hypothesis cannot be supported as there is no significant difference in academic performance between ability- and randomly grouped sections. The ability-grouped sections actually performed slightly worse overall on average (81.04% vs. 80.60%). While these results may have minimal impact on other undergraduate institutions that do not have the structure to offer academically homogeneous classrooms, it should be freeing to those school administrators and leaders that wish they could offer such an academic environment. In the end, the individual ability of the student is likely what drives their performance and not the ability of those around them. Perhaps the common strategy and perceived benefit of having higher ability classes should be reconsidered.

ORCID

Dusty Turner http://orcid.org/0000-0003-2998-8799

REFERENCES

1. Carbonaro, W. 2005. Tracking, students' effort, and academic achievement. *Society of Education*. 78: 27–49.
2. Hickman, R. 2007. *Ability Group Sectioning in an Undergraduate Calculus Curriculum*. West Point, NY: Center for Teaching Excellence at the United States Military Academy.
3. Kuhn, M. 2018. CARET Classification and Regression Training. R package version 6.0-79.
4. Loveless, T. 1998. *Making Sense of the Tracking and Ability Group Debate*. Washington, DC: Fordham Institute Publications.
5. Sagi, O., and L. Rokach. 2018. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8(4): e1249.
6. Slavin, R. E. 1987. Ability grouping in elementary schools: Do we really know nothing until we know everything? *Review of Educational Research*. 57: 347–350.
7. Viar, W. 2008. *Tracking and Ability Grouping*. West Point, NY: Center for Teaching Excellence at the United States Military Academy.

BIOGRAPHICAL SKETCHES

MAJ Dusty Turner was formerly an assistant professor and course director of MA256: Advanced Introduction to Probability and Statistics at the United States Military at West Point. He received his undergraduate degree in Operations Research in 2007 from West Point, a masters degree in Engineering Management from the University of Missouri at Rolla in 2012 and a masters degree in Integrated Systems Engineering at The Ohio State University in 2016. He currently serves as an Operations Research Systems Analyst for the Center for Army Analysis in Fort Belvoir, Virginia.

MAJ Jim Pleuss is an assistant professor and course director for MA206: Introduction to Probability and Statistics at the United States Military Academy at West Point. He received his undergraduate degree in Computer Science from West Point in 2007 and a masters degree in Operations Research from Kansas State University in 2016. He is currently an Operations Research Systems Analyst in the United States Army.

MAJ Christopher Collins is an instructor of MA206: Introduction to Probability and Statistics at the United States Military at West Point. He received his undergraduate degree in Engineering Psychology in 2003 from West Point, a masters degree in Engineering Management from the University of Missouri at Rolla in 2008 and a masters degree in Operations Research at Kansas State University in 2015. He currently serves as an Operations Research Systems Analysts in the United States Army.