Title
Hierarchical Linear Modeling Approach to Measuring the Effect of Class Size and Other Characteristics on Student Grades in Introductory Physics Courses

Author
Gorman, Connor Thomas

Publication Date
2022

Hierarchical Linear Modeling Approach to Measuring the Effect of Class Size and Other Characteristics on Student Grades in Introductory Physics Courses

By

CONNOR GORMAN
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Physics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

David Webb, Chair

Richard Scalettar

Jacob Hibel

Committee in Charge

2022

Dedication

This dissertation is dedicated to the struggle for full liberation, especially within, and often against, the University of California as it is currently structured. We must continue working to dismantle all systems of oppression (and their intersections) while simultaneously building a just and democratic world from the ashes of the old.

Table of Contents

Acknowledgments

First, I would like to acknowledge my advisor, Dr. David Webb, who brought up a variety of points that have been incorporated into this dissertation, both in terms of content and format, as well as for showing patience around my extended timeline. Relatedly, I would like to acknowledge the support and feedback I have received throughout my time studying Physics Education Research from the rest of the (sadly, soon to be gone) UC Davis PER group, especially from my friends, co-workers, and fellow graduate students Mary Chessey and Rylai Luginbuhl.

In addition, I would like to acknowledge my other committee members, Dr. Jacob Hibel and Dr. Richard Scalettar, the former of whom also taught two sociology courses that I took on (respectively) education and regression analysis, both of which ended up being quite relevant to this dissertation. Similarly, I would like to acknowledge Dr. Kevin Gee, who taught an education course on HLM where I first learned about this technique and realized how it could be applied to my research interests.

I would also like to acknowledge Angela Sharma, the UCD physics Graduate Program Coordinator during my entire time in the UCD physics department, whose time here almost exactly overlaps with mine. She was instrumental in helping me (and numerous others) navigate the bureaucracy of the department, grad studies, and the general UC administration. Next, I would like to acknowledge my parents, Carolyn and Tom Gorman, for their support and patience over the years.

Finally, I want to acknowledge all of my friends and co-workers who have provided support and helped guide me on a personal level, along with helping to shape my understanding of the world by teaching me so much about organizing, solidarity, and direct action. In particular, I would especially like to acknowledge the radical, liberation-oriented graduate students in UAW 2865, the UC Student-Workers Union. In fact, UAW 2865's demands around class size during our 2013/2014 contract campaign was a major inspiration for the primary focus of this dissertation.

Within the UAW, I would like to acknowledge the Anti-Oppression Committee (AOC), as well as the AWDU (Academic Workers for a Democratic Union), CLEW (Collective Liberation for Education Workers), and UAWD (Unite All Workers for Democracy) caucuses. These amazing groups fought, and in some cases continue to fight, for the intersecting needs of Academic Student Employees (ASEs), UAW members, and the wider community by challenging and resisting conservative tendencies within both the UC administration and UAW administration caucuses.

Abstract

The effect of class size on student learning has numerous policy implications and has been a major subject of conversation and research for decades. Despite this, few studies have been done on class size in the context of university settings or physics courses. This dissertation helps address that gap in the literature by quantitatively analyzing the effect of class size on students' understanding of physics concepts in active-learning based introductory physics courses for bioscience majors at a large, R1 university. In the process, this dissertation also discusses the reasoning and methods behind three-level basic Hierarchical Linear Modeling (HLM), which is a particular form of statistical regression, along with analyzing the effects of several additional, non-class-size related, student and class-level characteristics on student understanding of physics concepts.

In this study, a student taking a given course is part of a section which is itself part of a larger Lecture. It was found that Grades, which were used as a proxy for students' understanding of physics, varied a lot between individual students and also between Lectures, but varied relatively little between different sections within the same Lecture. Furthermore, these Grades were affected substantially by the student-level factors which were part of this study, including academic factors like GPA and repeating a course, as well as demographic factors like race and ethnicity. These Grades were also impacted greatly by Lecture-level factors that were part of this study, like academic term and Lecture instructor. However, these Grades were not consistently impacted by any of the section-level factors that were part of this study, including start times and mean GPA.

On class size specifically, within the ranges studied here, class size did not have much of an effect on Grades in the courses that were part of this study. There were some signs that larger Lecture sizes lead to lower Grades, but there was not enough evidence to be definitive, and there were no consistent trends in the impact that section size had on Grades. However, the relatively low variation in Grades between sections, together with a variety of other questions, issues, and limitations, means that this study is certainly not the end of the story. In particular, there are still open questions around the nature and meaning of Grades, as well as how class size and other student and class-level characteristics impact non-Grade related aspects of student, as well as teacher, well-being and success. There is still much work to be done and plenty of discussions to be had, both theoretically and empirically, when it comes to the types of courses that were part of this study and the factors that affect student understanding of the material covered by these courses.

Chapter 1: Introduction and Background

Class Size Motivation, Background, and Research Question

The effect of class size on student learning (sometimes referred to as achievement or understanding) has been a hotly contested and heavily researched topic for quite some time now. Studies show that when implemented properly, classes of 15-17 students are optimal, at least in the early grades (often Kindergarten through third grade) [1]. However, there is also evidence (using Multilevel Modeling) to suggest that the effect of class size may actually vary depending on other factors [2]. This implies that more research is necessary to determine the conditions under which class size is relevant and the different ways that its effects manifest themselves depending on the circumstances. It would therefore be beneficial to conduct such research in educational settings that have not yet been studied in the context of class size. One of these areas is secondary and post-secondary education and another one is physics courses, which have both had relatively few studies done on the issue of class size, though there are a few exceptions yielding mixed results [3, 4, and 5]. This means that additional class size studies in higher education and physics could be of use to the education community, both to progress the knowledge of student learning in these settings and to help inform where educators should focus their limited attention and resources.

Furthermore, physics education is moving in the direction of taking a more active approach to learning [6] and even with technological advancements, it is likely that for active-learning techniques to work, teachers will still have to be physically present [7]. It also seems reasonable to speculate that class size would have an even greater impact in an active-learning environment than it would in a more traditional environment since when students passively watch someone lecture, the number of students present is less likely to impact a given student's understanding of the material than it would if students are working together on problems with guidance from the instructor. Essentially, the number of students present has a higher chance of affecting classroom dynamics in the former case than the latter. Therefore, it is especially desirable to study the effect of class size in the context of physics courses that employ active-learning techniques. Quantitatively, this speculation would lead to measures of active-learning moderating the effect of class size on student understanding, though at least one study has found no significant interaction between a school's average class size and certain measures of active-learning when it comes to overall science achievement [8]. However, overall science achievement in that study was measured by a single standardized test in secondary school and is therefore not necessarily indicative of class size's effect on student understanding of physics material in an introductory college course.

Given all of the above, the main research question in this study is: "What is the relationship between class size and students' understanding of physics concepts in active-learning based introductory physics courses?" While the primary focus here is on class size, other factors, like student demographics, class start times, and Lecture instructors, were also included in this study's analyses as controls, thereby providing an opportunity to examine the effect of these factors on students' understanding of physics concepts in active-learning based introductory physics courses as well.

Statistical Motivation and Background

Physics Education Research (PER) is a rapidly growing field which exists in the physics community (while drawing from, and working with, researchers in other fields, like education and sociology) that seeks to determine and promote best practices around teaching physics. In order to make these determinations, it becomes necessary to compare and evaluate different teaching styles, techniques, and methods. There are a variety of ways to make such comparisons, both quantitatively and qualitatively, depending on the situation. However, since PER is still a relatively new field, many of these potential analyses have not been explored in a PER context even if they are frequently applied to other forms of education research.

Multilevel Modeling

One mechanism that is often used to evaluate the effectiveness of teaching interventions, and yet has not been widely adopted in PER thus far (though with a few notable exceptions, like some of the analysis in [9]), is Multilevel Modeling. The basic motivation behind Multilevel Modeling is that many situations exist where data points should be grouped together (clustered) as a result of the larger structures that they naturally form. One common example of this is the fundamental relationship between individuals and institutions, which underlies the social sciences, where individuals both influence, and are influenced by, the society and social structures that they are a part of [10, p. 1]. In this case, individuals are naturally clustered within institutions (schools, medical facilities, etc.) and geographic or geopolitical (municipalities, counties, countries, etc.) regions [11, p. 6]. Another common example involves longitudinal [10, p. 1] or repeated measures [12] research where observations, such as scores on a test or survey responses, are recorded multiple times, albeit on different occasions, for the same individual making these observations clustered within individuals. Two other interesting applications of Multilevel Modeling are meta-analyses where observations are clustered within studies [10, p. 1] and studies that involve multiple interviewers where interviewees are clustered within interviewers [11, p. xx].

For a variety of conceptual and statistical reasons, these types of situations warrant their own treatment rather than lending themselves to standard techniques. Conceptually, interpretations should always be made at the proper level where here “level” refers to either the smallest units being analyzed (the “lowest level” which frequently, though not always, means individuals) or one of the clusters that these units are grouped into. Trying to interpret something at any other level (for instance, trying to draw conclusions at the level of individuals based on an analysis that was done at the level of geographic regions, or vice versa) can lead to fallacies and inaccurate conclusions [10, p. 3-4]. Before delving into the statistical reasons for needing new techniques to address these types of situations, it should be mentioned that the underlying assumption behind Multilevel Modeling is that at each level, a random sample is chosen from the larger population (for instance, if the study involves students in schools, a random sample of schools is chosen and then a random sample of students is chosen from each school) [10, p. 1].

The main statistical reason why multilevel data (data that contains natural clustering) requires its own techniques has to do with observations’ independence, or lack thereof. Oftentimes, when standard techniques are applied to multilevel data, higher level observations (observations about the clusters themselves) are disaggregated, or in other words, assigned to a lower level like, though not necessarily, the lowest one [10, p. 2-3]. For instance, a school’s characteristics might be assigned to each individual student who attends that school. The problem here is that standard statistical techniques treat these lower level observations as independent when they clearly are not. Even beyond higher level traits' explicit lack of independence when they are disaggregated to lower levels, lower level observations that are naturally clustered (for instance, individual characteristics associated with students who attend the same school) will generally not be independent either, whether because of explicit relationships (for instance, all of the students at a given school being subject to the same grading policies which impact their individual grades) or because of implicit similarities (for instance, students at a given school coming from similar socioeconomic backgrounds) [10, p. 4-5]. Incorrectly treating correlated observations as independent leads to standard errors that are deceptively, and inaccurately, small [10, p. 4-5] (and thus, to deceptively and inaccurately narrow confidence intervals) which in turn, leads to results that appear more significant than they actually are [10, p. 3]. On the other hand, aggregating lower level observations by assigning some combination of them to a higher level (for instance, averaging student grades in each of several schools and then assigning these averages to their respective schools) eliminates information and therefore, reduces statistical power (for instance, an effect that is significant and would otherwise be detectable may no longer be) [10, p. 3].

The fundamental statistical problems in both of the scenarios described above revolve around a combination of non-independence and different sample, as well as population, sizes at different levels. Some very useful techniques have been developed over the years to deal with clustered and non-independent data by determining such things as effective sample sizes and effective variances [10, p. 5-6]. Effective quantities are mathematical tools that are designed to serve some computational or other practical purpose, like more accurately modeling a situation with non-independent observations in the case of effective sample sizes and variances. These quantities are generally different than their actual counterparts, like the actual number of observations in the case of sample size and the calculated variance (which is based on an assumption of independence) in the case of variance. However, by definition, Multilevel Models inherently contain multiple sample and population sizes (for example, there will always be more entities, like students, at the lowest level than entities, like schools, at any higher level) [10, p. 6]. Essentially, in Multilevel Models it is not simply the case that observations are not independent. Instead, on top of observations not being independent because of clustering, the actual sample and population sizes (i.e. number of entities) at different levels are different. Additionally, Multilevel Models often include variables that naturally exist at different levels (for instance, a model might contain variables that are associated with students along with variables that are associated with schools) [10, p. 6-7]. Therefore, it makes sense to develop and use new statistical techniques when analyzing multilevel data rather than trying to apply standard techniques with some slight modifications (like effective sample sizes and variances).

Hierarchical Linear Modeling (HLM)

Multilevel Models derive their name from the way different groupings or clusters in the data constitute different levels. For instance, if several observations are made on students in a few different classrooms at multiple schools during different years, the observations, students, classrooms, schools, and years all constitute different levels. A special type of situation occurs when these levels are sequentially nested inside one another, forming a hierarchy (for instance, observations that are nested inside students who are nested inside classrooms which are nested inside schools). However, this is not always the case with multilevel data. For example, in the above situation it is unclear whether schools or years would constitute a “higher level” or in other words, which one (if any) is nested inside the other [12]. Despite having more restrictions than general multilevel data, hierarchical situations with a strict nesting structure are relatively common and thus, it is useful to develop and apply techniques that are specific to these situations. The name that is often given to such analyses is Hierarchical Linear Modeling (HLM). In the case of nested data, HLM considers each of the sequential groupings to be a new level, where lower levels are nested within higher levels, and accounts for this nesting structure by associating each variable (along with the corresponding observations) with the appropriate level [11, p. xxi].

One of the most straightforward examples of this occurs in formal schooling where, as mentioned above, students are nested within classes which are themselves nested within schools or universities, as depicted in Figure 1 [11, p. xix]. In this example, level 1 would be the student level meaning all of the variables and observations describing students' individual traits (such as their gender or prior test scores) would be associated with level 1. Level 2 would be the class level meaning any variables and observations that have to do with the class a given student is in (such as the teacher's experience or how many students are in the class, i.e. the class size) would be associated with level 2. Finally, level 3 would be the school or university level meaning any variables and observations that have to do with the school or university a given student is in (such as median parental income or attendance policies) would be associated with level 3. In future sections, this example will be referred to as the Students/Classes/Schools example.

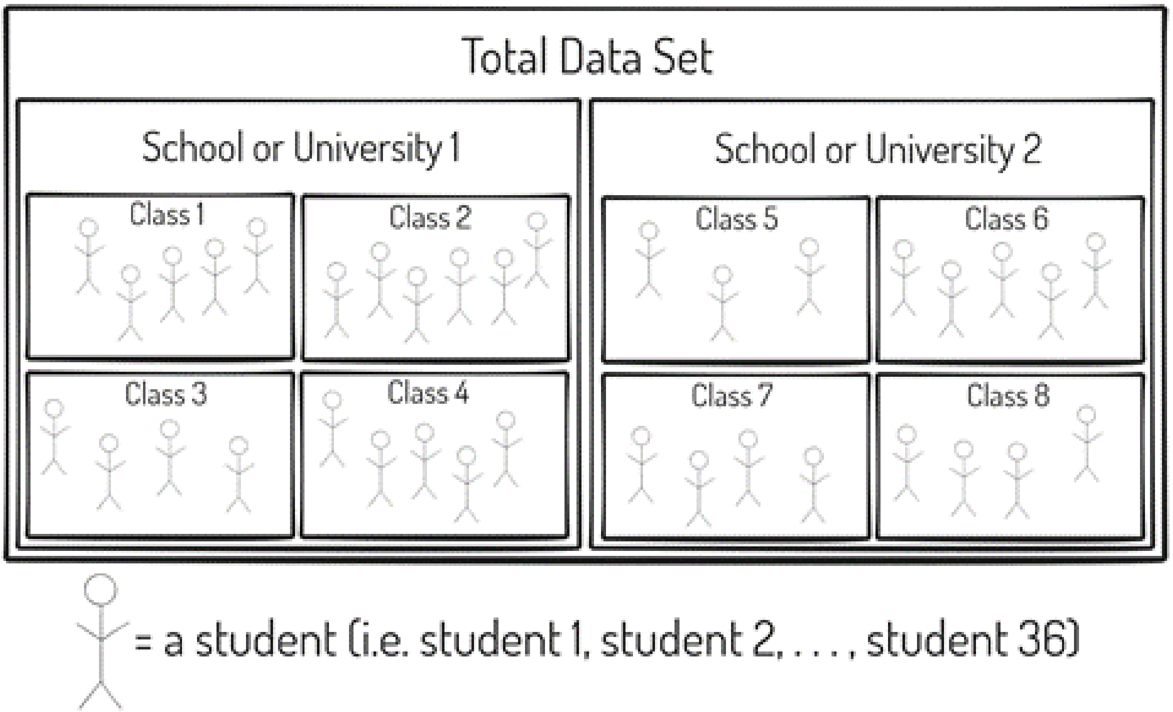

Figure 1: Illustration of the HLM nesting structure in the Students/Classes/Schools example.

Statistical Techniques and Random Effects

Multilevel Modeling

While all regression analyses involve quantitatively explaining the variation in some outcome (dependent) variable(s), in situations that involve multilevel data and that are analyzed using Multilevel Modeling, the mathematical distinction compared to standard multivariate regression techniques is the inclusion of random effects in addition to fixed effects [12]. Essentially, one general way that researchers can characterize the sources of variance (a particular type of variation that is common in regression analyses) in a given outcome variable is by grouping these sources into three categories; 1) things the researchers can measure and are interested in (because these things help answer the research question(s)), 2) things the researchers are not interested in but can measure and would like to control for (because of their likely impact on the analyses' results for things that the researchers are interested in), and 3) things the researchers are not interested in and cannot measure but would still like to control for (again, because of their likely impact on the analyses' results for things that the researchers are interested in) [12].

Starting to get into the statistical details, there are three broad ways of doing a regression analysis on multilevel data; a single standard regression (like in Figure 2), separate standard regressions that are each applied to a different cluster (like in Figure 3), or Multilevel Modeling (like in Figures 4 and 5). In all of these, the first two sources of variance described above are accounted for using what is known as fixed effects where predictor (independent) variables are multiplied by coefficients (often referred to as slopes) in a regression equation. Each of these coefficients, together with that coefficient's standard deviation, tells the researchers about the corresponding predictor variable's effect on the value of the outcome variable (though other information is often required to determine how much a given predictor variable explains the variance in the outcome variable).

However, it is extremely rare that researchers can determine and include all relevant predictor variables or that predictor variables can fully explain all of the variance in the outcome variable. Therefore, the third source of variance mentioned above leads to unaccounted for variance in the outcome variable, as well as related error terms (the difference between actual values of the outcome variable and predicted values of the outcome variable based on the regression equations). A portion of this variance, known as process variance, is due to unaccounted for predictor variables and can be addressed through random effects, which essentially means having a regression analysis that allows the outcome variable's intercept and/or the slopes on one or more of the predictor variables to be different for different clusters [12]. Furthermore, as discussed in "Statistical Motivation and Background," not accounting for natural clustering would violate the regression assumption that observations are independent. Not accounting for such clustering would also violate the regression assumption of homoscedasticity, which says that the variance in error terms is the same across all values of each predictor variable [11, p. xx]. Incorporating random effects is an effective way of accounting for such clustering while still maintaining as much statistical power (the ability to determine statistical significance) as possible, which would not be the case if a separate standard regression were applied to each individual cluster (due to drastically reduced sample size).

For instance, expanding on the Students/Classes/Schools example, say the outcome variable is a student's score on some exam. In this example, holding all of the predictor variables at particular values leads to a particular value of the outcome variable (i.e. a particular exam score) but if the regression analysis allows the value of the outcome variable (when all of the predictor variables are held at particular values) to differ between classes and/or schools or universities (rather than assuming it is the same across classes and schools or universities), then this would be a situation with random intercepts.

Now, if in the Students/Classes/Schools example a student-based predictor variable, say a student's average grade on prior exams, is part of the regression analysis, then there will be a regression coefficient (slope) associated with this predictor variable that tells researchers how much a given student's exam score changes (on average when all other predictor variables are held constant) when their average grade on prior exams changes by one unit (in whatever units grades are being measured in). If the regression analysis allows the value of this slope to differ between classrooms and/or schools, then this would be a situation with random slopes. Similarly, if a class-based predictor variable, say the teacher's temporal experience in years, is part of the regression analysis then there will be a slope associated with this predictor variable that tells researchers how much a given student's exam score changes (on average when all other predictor variables are held constant) when their teacher's experience changes by 1 year. If the regression analysis allows the value of this slope to differ between schools or universities, then this would also be a situation with random slopes. Note that the Students/Classes/Schools example is a case of HLM where students are not only clustered into both classes and schools or universities, but classes are themselves clustered into schools or universities (i.e. classes are nested within schools or universities). However, the discussion here still applies to situations where this nesting structure does not exist, though there are some important aspects of HLM that do not apply to all Multilevel Modeling situations.

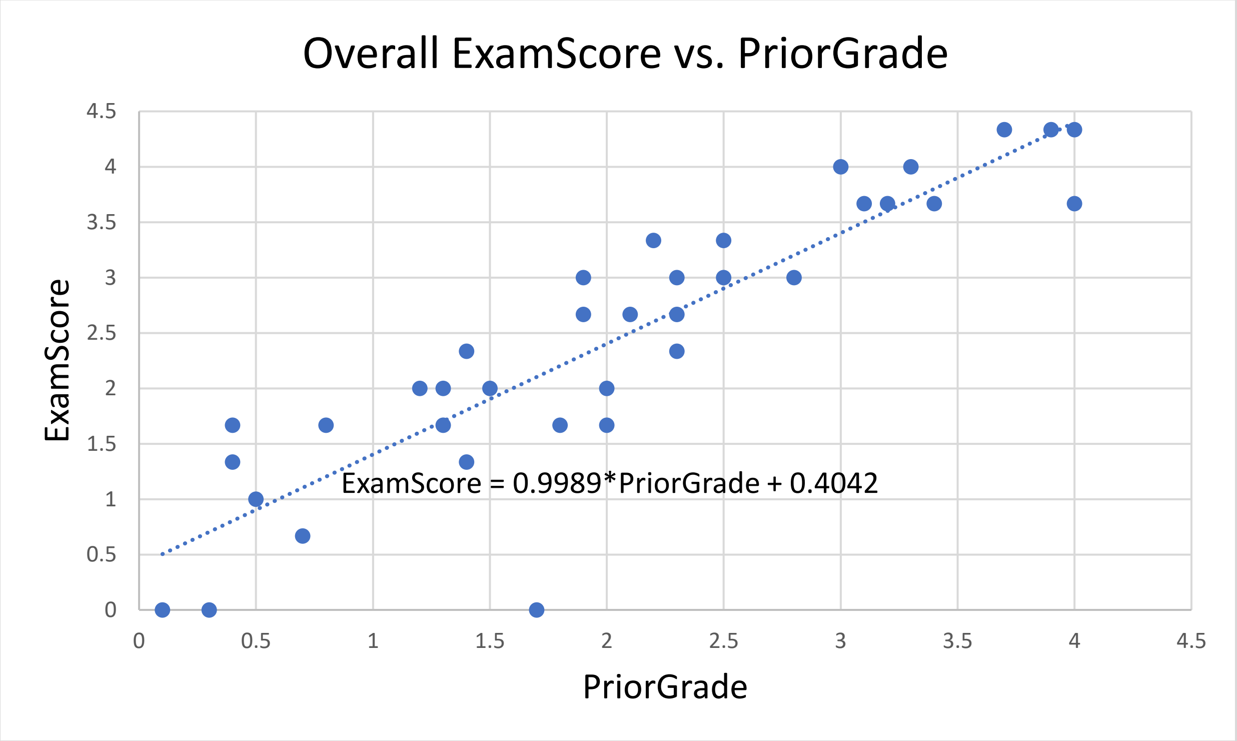

Figure 2: A standard regression plot, using example data, of student scores on a particular exam (ExamScore) vs. their average grades on previous exams (PriorGrade) in the Students/Classes/Schools example.

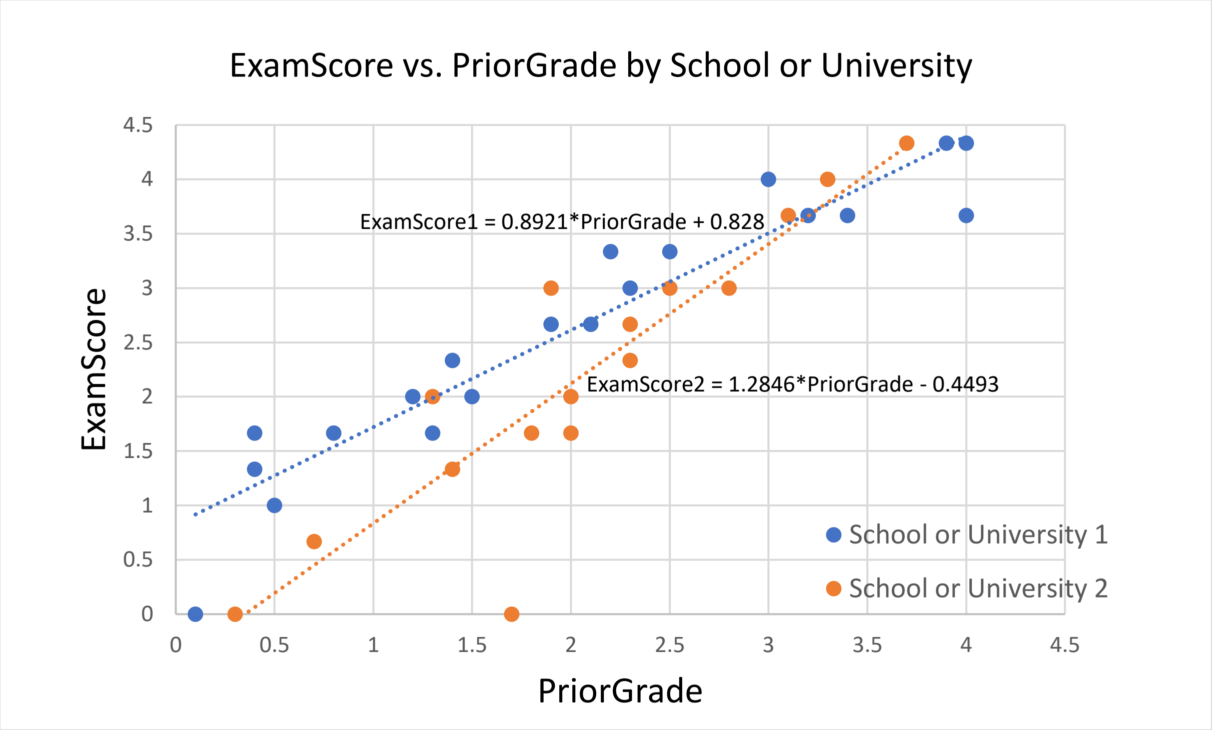

Figure 3: A plot (using the same example data as in Figure 2) of ExamScore vs. PriorGrade at two different schools or universities in the Students/Classes/Schools example. Note that both the intercepts and the slopes associated with each school or university are different compared to both each other and their respective values in Figure 2.

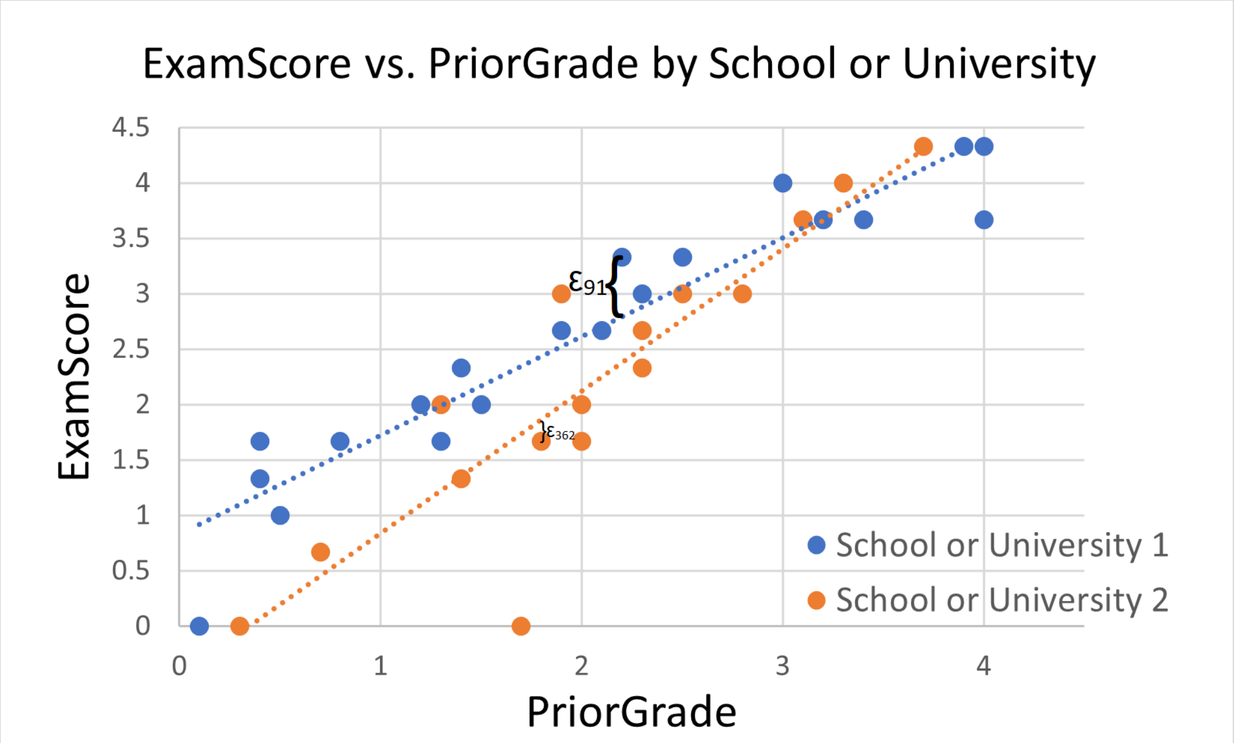

Figure 4: The same plot as in Figure 3, but with an illustration of residual error terms; in this case the difference between a given student's actual ExamScore (outcome variable) value and that which is predicted by their PriorGrade (predictor variable) value for the school or university (cluster) that they are part of. ε_{91} *is the residual error term for student 9 in school or university 1 while* ε_{362} *is the residual error term for student 36 in school or university 2 (*ε_{362} *on Figure 4 is a bit small, but is the only other labeled error term in Figure 4 besides* ε_{91} *and it is the only labeled error term in Figure 4 for school or university 2). More details can be found in Appendix A, though note that this illustration is more akin to a two level model than a three level model (which is why* ε *does not include a number for the class that a given student is in), but the basic ideas here are the same.*

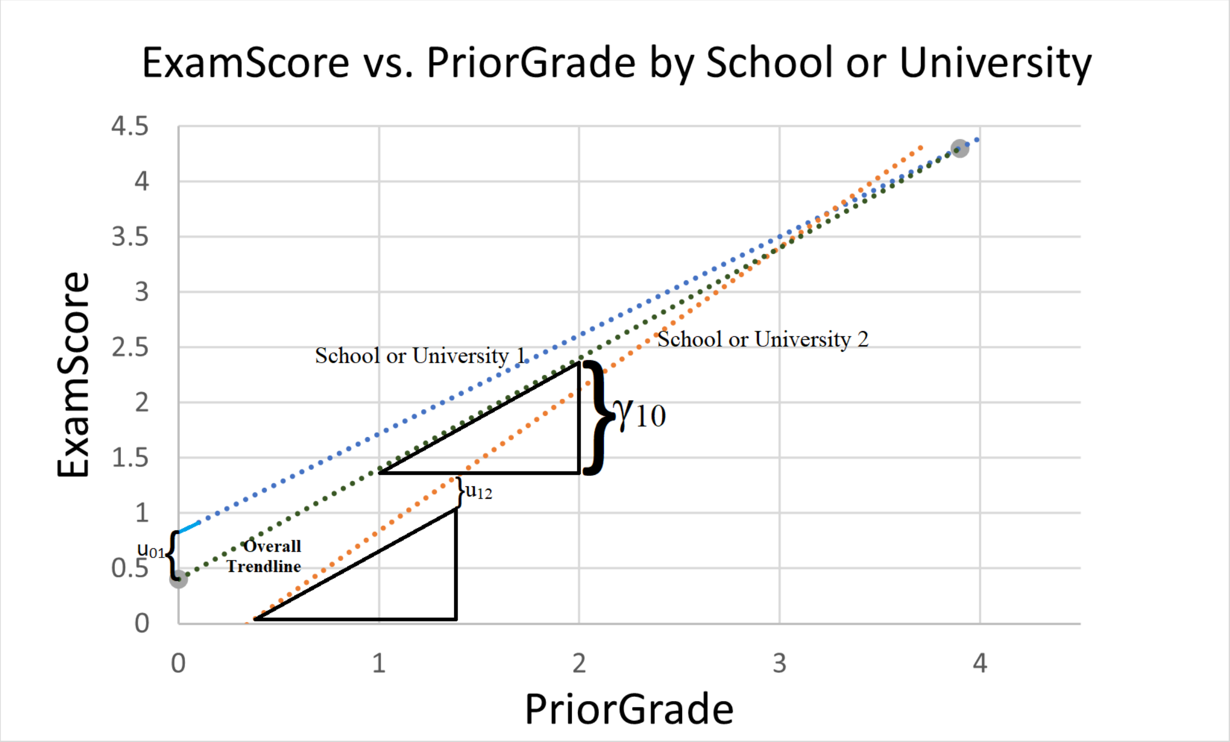

Figure 5: A plot with the same axes and scaling as Figure 3, but with the trendlines from both Figure 2 and Figure 3 superimposed onto each other, while removing the actual data points in order to focus on the trendlines. u_{01} is an error term that is the difference between the intercept of the overall trendline (in green, from Figure 2) and the intercept of the trendline for school or university 1 (in blue, from Figure 3). γ_{10} is the slope of the overall trendline (the increase in outcome variable ExamScore due to a 1 unit increase in predictor variable PriorGrade based on all of the example data) while u_{12} is the difference between the slope of the overall trendline and the slope of the trendline for school or university 2 (in orange, from Figure 3). More details can be found in Appendix A, though note that this illustration is more akin to a two level model than a three level model, but the basic ideas here are the same.

HLM

As described in “Statistical Motivation and Background,” HLM is a type of Multilevel Modeling that imposes additional restrictions, where levels are sequentially nested inside of one another. Despite having more limited applicability, hierarchical situations are relatively common, especially within education settings, and also lend themselves to more systematic approaches and interpretations. In terms of numerical results, HLM is not different from any other type of Multilevel Modeling, but conceptually, there are interesting and useful features of HLM that rely on its nesting structure.

The main reason for this is that intercepts and slopes on predictor variables are not only allowed to differ between different clusters, but can themselves be modeled and interpreted as functions of predictor variables (specifically, predictor variables associated with the next highest level as detailed in Appendix A). This makes keeping track of a large number of predictor variables and coefficients more manageable and also naturally leads to an interpretation where lower level intercepts and slopes on predictor variables are partially explained through higher level predictor variables (along with error terms which form the mathematical basis for random effects). Furthermore, it facilitates an iterative regression process where predictor variables can be introduced one level at a time, thereby helping researchers determine the amount of variance in the outcome variable that is explained by each level's predictor variables. Note that in this method of analysis and interpretation, neither intercepts nor slopes can be directly related to predictor variables more than one level above them, but they can be indirectly linked to such predictor variables through higher level equations.

For instance, going back to the Students/Classes/Schools example, where now the nesting structure is being accounted for, if the outcome variable (exam score) has different intercepts (values of the outcome variable when all predictor variables are held at particular values) for different classrooms then these intercepts may be partially explained by the values of classroom level predictor variables. Similarly, if the slope associated with the level 1 (student level) predictor variable of average grade on prior exams is different between different classrooms then this slope may also be partially explained by the values of classroom level predictor variables. In this example, neither the intercept nor the slope mentioned above can be directly related to level 3 (school level) predictor variables, but they can be indirectly linked to level 3 predictor variables through higher level equations (i.e. level 2 coefficients being treated as functions of level 3 predictor variables).

While there are more advanced HLM techniques, the most basic version requires a strict nesting structure and for the outcome variable to be continuous (at least interval, if not ratio, as opposed to categorical or ordinal), at level 1 (which is often, though not always, the level of individuals), and to follow an approximately Normal distribution. Note that even though it is theoretically possible to have any (whole) number of levels, in practice HLM analyses often only include two or three levels since having more than three levels can be computationally difficult (i.e. lead to convergence issues) and can also be complicated to interpret [10, p. 32-33]. In general, the more levels are added, the more reliant researchers need to be on theoretical assumptions that simplify the mathematical model [10, p. 33]. Check out Appendix A for more details on the technicalities and equations involved in the three level version of a basic HLM analysis.

Random Effects and Levels vs. Predictor Variables

One final issue to mention here is the distinction between random effects and sets of categorical predictor variables (sets of dummy variables where every observation is assigned a value of 1 for exactly one such variable, typically the one associated with the name of the variable, and 0 for all of the others). In principle, the clusters in any Multilevel Modeling analysis (including any HLM analysis) could be treated as a set of categorical predictor variables (i.e. using fixed effects). For instance, in the Students/Classes/Schools example, it would be possible to use a conventional regression analysis which includes a categorical variable for each class and a categorical variable for each school or university such that each of these variables has a value of 1 for any students who are in that class or that school or university and 0 for all students who are not. A reference class and a reference school or university would then be chosen and the corresponding variables would be excluded from the analysis while all other such variables would be included (with their corresponding slopes comparing them to the reference class for class variables and the reference school or university for school or university variables). This type of analysis would also allow for interactions between these variables and any or all other variables, such as average grade on prior exams, teacher experience, and a school's median parental income.

However, when natural clustering is involved and the analysis does not require a comparison of outcome variable means between different clusters, accounting for this clustering using random effects creates stronger statistical power, especially when there are a large number of clusters [12]. Additionally, in the case of HLM, the conceptual benefits of doing such an analysis (mentioned previously) are another reason to account for this clustering using random effects and nested levels as opposed to categorical predictor variables.

A more mathematical way to determine whether to treat clustering using random effects or categorical predictor variables is by calculating intraclass correlation coefficients (ICCs) for the situation where no predictor variables are present. ICCs give the proportion of total variance in the outcome variable that exists between clusters (for instance, in the Students/Classes/Schools example this would mean the variance between classes or between schools or universities), as opposed to within clusters (for instance, the variance between students within a given class in the Students/Classes/Schools example). Check out the end of Appendix A for details.

Chapter 2: Study Set-Up

Site, Sample, and Population

The observations used in this study (the sample) were drawn from five years (2012 – 2016) of data (the period of study) on student Grades as well as student, discussion/lab, and lecture characteristics in a sequence of three introductory physics courses (7A, 7B, and 7C) designed for, and taken primarily by, bioscience majors at a large, public, R1 university that is on the quarter system (the University). All three of these courses are offered during all three quarters (Fall, Winter, and Spring) of the regular academic year, as well as during both of the University's two yearly summer sessions (one during the earlier part of summer break and one during the later part). The data used in this study was acquired from the University administration by a professor at the University for the purposes of education research.

The three courses that were part of this study are based on active-learning techniques where, during the regular academic year, students attend two discussion-labs (DLs) of 2hr and 20min each per week. In DL, students engage in activities like working through practice problems (similar to discussion sections in more traditional courses) and conducting and analyzing experiments (similar to lab sections in more traditional courses) in groups that are typically composed of five students each, though the way that these groups are determined differs between DLs, with some being chosen randomly and others being self-selected. During DL, students also synthesize material as a whole class. Both of these components (small-group activities and whole-class discussions) are done with the guidance of a teaching assistant (TA). The main DL classrooms have room for six groups which means 30 students when there are 5 students per group, though this number (and thus, the number per group) can vary. There is also an overflow classroom that is sometimes used and has room for four groups.

During the regular academic year, students also attend a single weekly lecture section of 1hr and 20min which typically provides an overview of the material, goes over some example problems, and is where exams and (often weekly) quizzes are administered. Lecture instructors for these courses teach two identical lectures back-to-back with students from both of these lecture sections being part of the same DL (i.e. DLs consist of an approximately even mixture of students from two different lecture sections that are taught consecutively by the same Lecture instructor). It is important to note that, while most class-time and class-based learning for the courses that were part of this study is done during DL, and Grades for these courses do include some slight modifications for DL participation (which tend to be approximately equal on a per DL basis and therefore do not impact the relative Grade differences between DLs), these Grades are usually based almost exclusively on quiz and final exam (and sometimes midterm, which are rare but not unheard of) scores. With a few exceptions, quizzes and exams are written by the Lecture instructor and administered during lecture sections.

During the summer (in each of the two yearly summer sessions), the general layout and policies of these courses remains largely the same as during the regular academic year, but with a few key differences. First, only one lecture section of each course is offered in any given summer session, so Lecture instructors only teach one lecture section at a time and thus, all students in every DL come from the same lecture. Summer Lectures also tend to not be full, so the number of students per DL (and lecture section) is often lower than it is during the regular academic year. Furthermore, summer Lectures and DLs have a pace which is approximately double that of the regular academic year, meaning during the summer, there are four weekly DLs (which are still 2hr and 20min each) plus two weekly lectures (which are 1hr and 15min each, so 5min shorter than during the regular academic year). Finally, external factors, including the number of courses a given student is taking simultaneously, the weather (the University is in an area that gets quite hot and dry over the summer), and the types of students who choose to take these courses, are different during the summer than they are during the regular academic year.

Lastly, there is a question of what population the results of this study apply to (where this population is the larger group that the sample is assumed to be randomly drawn from). Here things become a bit more complicated because this question essentially means, how generalizable is this study? Do these results apply to all courses, all university courses, all physics courses, all university physics courses, etc.? Or perhaps they apply even more narrowly. For instance, maybe they only apply to active-learning based courses or active-learning based physics courses. Perhaps they are particular to bioscience majors or bioscience majors in active-learning based physics courses. Given all this uncertainty, it is best to list the populations that this study, and its results, definitely apply to and then provide some discussion around larger populations that it might apply to. The population that this study's results definitely applies to is students who take these particular courses at the University, both during and outside of the time period that was studied. Getting a bit broader, it seems likely that this study's results apply to active-learning based introductory physics courses taken by students who are not specializing in the physical sciences, math, or engineering. This is because even though the courses in this study were college courses as opposed to high school courses and were taken primarily by bioscience majors as opposed to a general audience or non-STEM majors, it does not seem likely that these particular details would have a substantial impact on the results. Lastly, it is possible that the results of this study apply to active-learning based introductory physics courses taken by students who are specializing in the physical sciences, math, or engineering. On the other hand, it does not seem likely that the results of this study would apply to upper level or graduate physics courses nor does it seem likely that they would apply to courses in fields beyond physics, with the possible exception of active-learning based introductory chemistry, math, or engineering courses, especially those taken by students who are not specializing in the physical sciences, math, or engineering (in other words, courses taken by a general audience or students who are specializing in bioscience or non-STEM fields).

Levels and Outcome Variable

This study involves HLM analyses that use three nested levels. In these analyses, individual students (or really, observations, but these essentially amount to students) are at level 1, DLs are at level 2 such that students are nested within DLs, and Lectures are at level 3 such that DLs are nested within Lectures. Note that, when it comes to data from the regular academic year, Lectures are defined to include both lecture sections that are taught back-to-back by the same instructor since these two lecture sections tend to be quite similar, albeit with some differences, including the time, how tired the Lecture instructor is, and how experienced the Lecture instructor is with teaching that day’s material. It would also be rather difficult to treat these two lecture sections as distinct while simultaneously maintaining the DL level (level 2) and a strict nesting structure (since most DLs during the regular academic year have students from both of these lecture sections). The different lecture sections within the same “Lecture” during the regular academic year are taken into account using a dummy variable (a predictor variable that is binary and has a value of either 0 or 1) as described in “Level Choices and Predictor Variables.” During the summer, Lectures and lecture sections are equivalent due to the way that lectures are structured during the summer, as described in “Site, Sample, and Population.”

The outcome variable in this study is Grade which represents the final numerical (on a 4.00 scale) grade for a given student in a given Lecture of a given course (7A, 7B, or 7C) during a given quarter of a given year (note that the term “course” is defined in this study as one of the three types of courses 7A, 7B, and 7C, as opposed to a particular Lecture or DL). There are 13 possible Grades, corresponding to 13 possible letter grades (including pluses and minuses), which range from 0 (F) to 4.33 (A+) in increments of 0.33 (or 0.34 in some cases). There are two main reasons for this choice of outcome variable. First, since this study’s main research question involves student understanding of physics concepts, the outcome variable should measure this understanding, which, at least in principle, final Grades supposedly do (check out Appendix B for more details about the research into what grades measure and the degree to which they actually measure what they are supposed to, along with some discussion of this study as it relates to such research). The second reason this outcome variable was chosen is because, regardless of their relationship to student understanding, in the context of our current society, grades are relevant in-and-of themselves since they are a common feature of most students’ educational experiences and are used as sorting or ranking mechanisms to determine who qualifies for certain benefits [13]. Essentially, reviewers often use grades to help determine which students (or former students) have access to resources and opportunities like jobs or additional levels of schooling (like internships, medical school, and graduate school). This is the primary reason that most students tend to be concerned about their grades and the courses involved in this study are no exception.

On the statistical side of things, note that Grade exists at level 1, is assumed to be continuous, and follows an approximately Normal distribution (as depicted in Figure 6), which allows analyses that use it as the outcome variable to be conducted with basic HLM techniques (provided the other conditions of a basic HLM analysis are met). The numerical Grade associated with an A+ letter grade was changed from its original value of 4.00 to a value of 4.33 in order to match the association between other letter grades with pluses and their corresponding numerical Grades, and also because many students still try to get an A+ even when they know that it will not affect their GPA any differently than an A would. Another benefit of this conversion is that it made the distribution of final numerical Grades closer to a Normal distribution than it would have been had all of the A+ Grades been assigned a value of 4.00, though there is still a ceiling effect and a disproportionately high number of 0s (Fs).

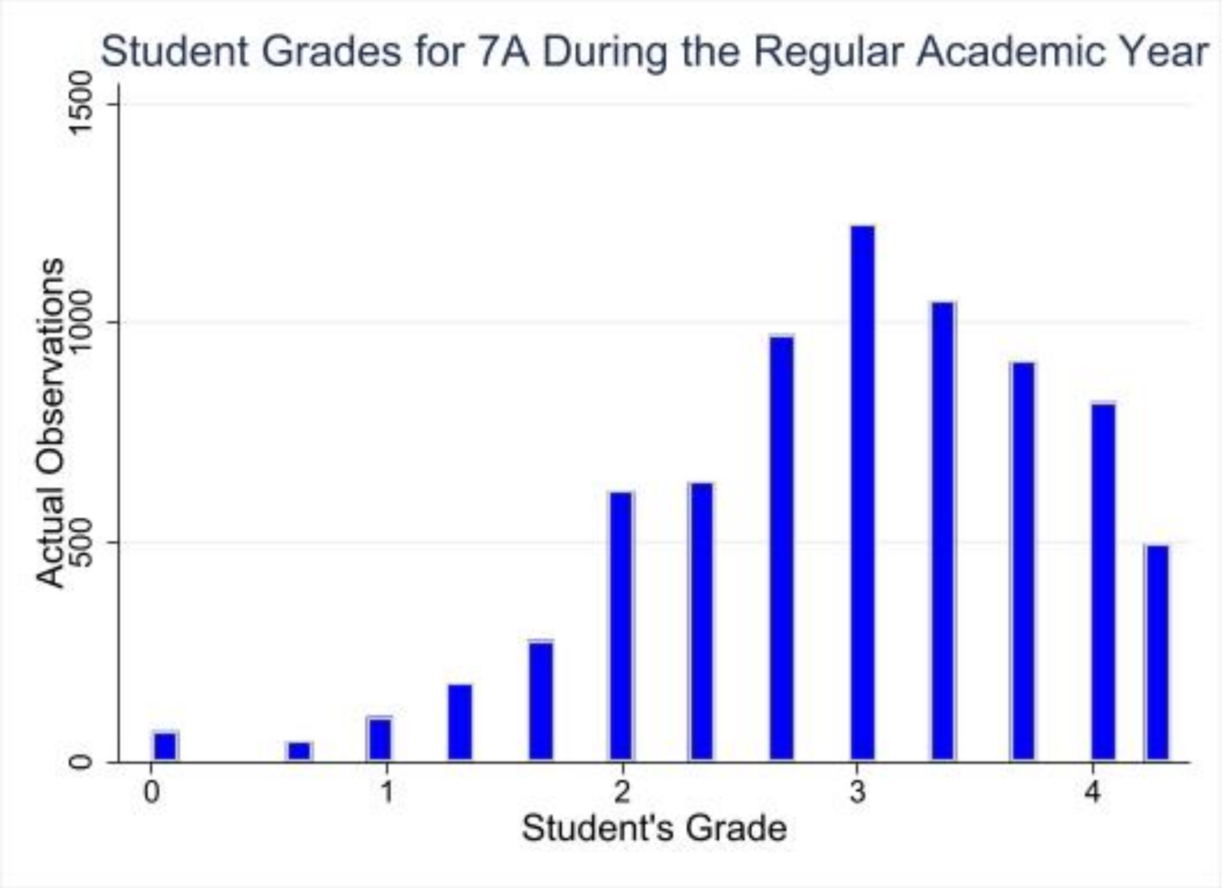

Figure 6: A histogram of Grades during the period of study for 7A during the regular academic year. This distribution is similar to the distribution of Grades in 7A during the Summer as well as in 7B and 7C during both the regular academic year and the Summer. The main deviation is in 7C during the summer where there is a slight dip at the peak. The peak also occurs at a slightly higher value of Grade for 7A and 7C during the summer.

The assumption of continuity (which also applies to the predictor variables GPA, Units, Mean_GPA, Mean_Units, Mean_Male, Mean_LecStart, and LecSize in this study) has two major components. First off, it means that the number of possible values is fairly large, but what "fairly large" entails is not well defined. For instance, both 13 values and 101 values (the number of integers on a 100 point scale) are discrete scales and thus, strictly speaking, neither of them is truly continuous in the uncountably infinite sense. The only difference between the two in this regard is how many distinct values (grades) are allowed and the amount by which consecutive values are separated and yet, while the legitimacy of treating letter grades as continuous is sometimes disputed, treating a 100 point scale as continuous is rarely questioned. It can be reasonably argued that 13 values are enough to qualify as continuous provided the second, arguably more important, criterion is met.

The second aspect of what it means to be continuous is directly tied to the question of value separation and whether the difference between consecutive values is meaningful. For ordinal variables, this is not assumed to be the case and consecutive values simply represent an ordered ranking without any meaning attached to the degree of separation. For continuous (interval or ratio) variables, it is assumed that the difference between consecutive values has the same meaning anywhere along the spectrum of allowed values. For instance, treating Grade as continuous here means that the difference between a C- (1.67) and a C (2.00) is assumed to be the same as the difference between a B (3.00) and a B+ (3.33), which is not necessarily the case. However, the same would be true on a 100 point scale where it is not necessarily the case that the difference between say, a 60 and a 61, is the same as the difference between an 84 and an 85, but in both cases this assumption needs to hold in order for numerical grades to be treated as continuous variables. Furthermore, in both cases it is reasonable to assume that it holds to good approximation (and if it does not, then there would be a lot of other philosophical and pedagogical problems that would need to be contended with).

On top of these potential issues with treating Grade as continuous (which have hopefully been addressed to the reader’s satisfaction), curves are sometimes implemented when it comes to grading introductory physics courses. In such a situation, one could argue that the corresponding grades are a ranking system within a given class (an ordinal variable) but do not present an interval type of knowledge since the difference between consecutive grades at one point along the grade spectrum does not necessarily have the same meaning as the difference between consecutive grades at a different point along this spectrum. Another way that grades are sometimes manipulated, which may raise concerns around the degree to which they can be treated as continuous, is shifting the thresholds or cutoffs for obtaining a given letter grade (which is a different type of correction for relative difficulty than a true curve even though this practice is sometimes referred to as curving). However, as far as the author is aware, the courses involved in this study are rarely curved and try to avoid large shifts in the thresholds for obtaining a given letter grade (compared to some conventional standard). In addition, since grading occurs within Lectures, which are being treated as a level in an HLM framework, there is already a built-in mechanism to account for possible differences in grading schemes between different Lectures. This means that in order for Grade to be continuous, it simply needs to be continuous within each Lecture rather than needing to be continuous across Lectures (i.e. while accounting for different grading patterns between Lectures), which is just another benefit of conducting the analyses in this study using Multilevel Modeling.

Level Choices and Predictor Variables

Before discussing the specific predictor variables in this study, it is informative to briefly discuss some of the choices that were made around the ways that different types of clustering were treated. In this study, DLs and/or Lectures could, in principle, be treated as sets of categorical predictor variables. However, given the large number of DLs and Lectures in the data set, along with the conceptual benefits of interpreting regression analyses using HLM when nested clustering exists, it is beneficial to treat these two types of clusters as (higher) levels within an HLM framework.

On the other hand, since there are a small number of quarters during the regular academic year (three; Fall, Winter, and Spring), it is reasonable to treat these as a set of categorical predictor variables rather than addressing them through random effects. Furthermore, treating these using random effects in a basic HLM framework would require either making academic quarter level 3, and not including Lectures in the analyses, or making it a fourth level that Lectures are nested within, neither of which would be desirable (the former because Lectures had a substantial impact on Grade, and the latter for the reasons discussed at the end of the "HLM" portion of "Statistical Techniques and Random Effects").

Academic year was not included in this study's analyses in any way (as a level or as a set of categorical predictor variables) since there is no reason to suspect that it would have any impact on Grades. Essentially, none of the years during the period of study have any defining characteristics that distinguish them from any of the other years (unlike, for instance, if the data had included observations from 2020 or 2021 when teaching and grading at the University changed substantially due to the COVID-19 pandemic). Furthermore, including academic year as a level would require either making it level 3, and not including Lectures in the analyses, or making it a fourth level that Lectures are nested within, neither of which would be desirable (because of the same reasons as those described above for academic quarters).

Lecture instructor was treated as a set of categorical predictor variables since, while there were quite a few of them (40) during the period of study, this number was still low enough to manage effectively using a set of categorical predictor variables (especially since not all of instructors taught each of the three courses in this study during the period of study) and including the Lecture instructor as a level would require either making it level 3 in place of Lectures, or making it a fourth level that Lectures are nested within, neither of which would be desirable (the former because, while the effect of Lectures on student Grades is partially due to the effect of Lecture instructors, the effect of Lectures is not necessarily limited to the effect of Lecture instructors).

Finally, note that each course in this study (7A, 7B, and 7C) was analyzed separately to remove the non-independence of students with themselves (i.e. the Grades received by a given student who took more than one of these courses will probably not be independent from one another). If the data was all analyzed together, rather than separated by course, the best way to account for this non-independence would be to treat observations as level 1 and students as level 2 (and DLs as level 3 and Lectures as level 4) such that observations are nested within students. Since this would add a fourth level and students would not be strictly nested within DLs, or even Lectures, under this structure (meaning it would not be possible to use basic HLM techniques), analyzing these courses separately is the most effective way to address this issue. Plus, treating each course separately also accounts for any unaddressed differences between the courses themselves. This could also be accomplished by making a course level (that Lectures are nested within) or a set of categorical predictor variables for the three courses in this study, but doing so would complicate the analysis in other ways (such as adding a fourth level if there was a course level or requiring a lot of interaction terms if the courses were treated as a set of categorical predictor variables).

Level 1 Predictor Variables

Binary Sex: Male, Female, and UnS are a set of categorical variables representing the binary sex that students identified with. A set of categorical variables is a set of dummy variables where every observation is assigned a value of 1 for exactly one such variable (typically the one that is associated with the name of the variable, as is the case throughout this study) and 0 for all of the other variables in the set. Sets of categorical predictor variables always have a reference category that the effects of the set's other variables/categories on the outcome variable are determined with respect to. The reference category here is Female since there are more female identifying students than male identifying students in the data (as shown by Table 2 in "Analysis Format and Data"). Female is 1 for female identifying students, 0 for male-identifying students, and 0 for observations where neither male nor female was listed in the data provided by the University. Male is 1 for male identifying students, 0 for female identifying students, and 0 for observations where neither male nor female was listed. UnS was defined by the author and refers to Unidentified Binary Sex. It is 1 for observations where neither male nor female was listed and 0 for both male and female identifying students.

Race and Ethnicity: AF, AI, CH, EI, FP, JA, KO, LA, MX, OA, OT, PI, VT, WH, and UnE are a set of categorical variables representing the race and ethnicity that students identified with where WH (White/Caucasian) is the reference category (due to both the demographics of the courses in this study and the structural barriers that People of Color experience at the University and within the wider society). Table 3 in "Analysis Format and Summary Data" includes the meaning of each category as defined by the University (with the exception of UnE, which refers to Unidentified Race and Ethnicity and was defined by the author as a category for observations where the race and ethnicity field provided by the University was blank).

U.S. Citizenship Status: Cit, PR, NI, RF, PO, IM, and UnC are a set of categorical variables representing students' U.S. citizenship status where Cit (U.S. citizen) is the reference category (due to both the demographics of the courses in this study and the structural barriers that non-U.S. citizens experience at the University and within the wider society). Table 4 in "Analysis Format and Data" includes the meaning of each category as defined by the University (with the exception of UnC, which refers to Unidentified U.S. Citizenship Status and was defined by the author as a category for observations where the U.S. citizenship status field provided by the University was blank).

Graduate Student Status: Grad is a dummy variable with the value 1 for graduate students and 0 for undergraduate students.

Course Repeats: Repeat is a dummy variable designed to account for the effect of some students repeating courses. For students who repeated a given course (7A, 7B, or 7C) for a letter grade at least once during the period of study, Repeat has a value of 1 for each observation associated with that student in that course, except for the observation associated with the first time they took that course, since this observation is not impacted by any prior experience with said course. Repeat has a value of 0 for all other observations (both observations associated with students who did not repeat a given course and those associated with students who did repeat a given course but where the observation in question is the first time they took that course during the period of study). Note that Repeat only accounts for repetition of a course for a letter grade and does not account for students who dropped a given course one or more times and then took it for a letter grade one or more times. This is because students drop courses for all sorts of reasons that are hard to quantitatively address and also because dropping courses usually occurs within the first few weeks of the quarter meaning a student's experience prior to dropping a given course should not have much of an effect on their understanding of the material, and thus, their Grade, if and when they take that course again.

Also note that Repeat only applies to retaking the same course for a letter grade meaning a student who took 7A, 7B, and 7C for letter grades during the period of study but never retook any of these three courses would have a Repeat value of 0 for all three of the observations associated with them in the data. As another example, a student who took (for letter grades) 7A, 7B twice, and then 7C during the period of study would have a Repeat value of 1 for the second (chronologically) 7B observation associated with them but a Repeat value of 0 for the other two observations associated with them (which correspond to taking 7A and 7C), as well as for the observation associated with the first time they took 7B. Similarly, a student who took 7A twice, 7B once, and then 7C three times during the period of study would have a Repeat value of 1 for the second 7A observation and the second two 7C observations associated with them but would have a Repeat value of 0 for the 7B observation, the first 7A observation, and the first 7C observation associated with them.

Because relatively few students retook these courses for a letter grade (as shown in Table 2), and most of those who did only retook a given course once, it is sufficient to treat Repeat as a dummy predictor variable as opposed to including an additional level for observations such that observations are nested within students (which would become level 2 if this were the case since level 1 would be the level of observations).

Lecture Start Time: LecStart is a dummy variable representing the start time of the lecture section that is associated with a given observation during the regular academic year. Because of the lecture structure described in "Site, Sample, and Population," during the regular academic year this variable not only accounts for lecture start times, but also accounts for other differences between the two lecture sections that are part of the same Lecture (i.e. that are taught back-to-back by the same Lecture instructor). This lecture structure is also why LecStart is at the student level during the regular academic year, as opposed to the Lecture level (a given Lecture during the regular academic year does not have a single start time due to the way that Lecture has been defined in this study, as described in "Levels and Outcome Variable"). LecStart is 0 for the earlier lecture time (7:30am during the regular academic year) and 1 for the later lecture time (9:00am during the regular academic year).

Originally, LecStart was included as a dummy predictor variable during the summer as well, though at level 3 since this makes more sense than treating it as a student level predictor variable if it is possible to treat it as a Lecture level one, and during the summer, it is (since, as mentioned above, during the summer a given Lecture consists of a single lecture section while during the regular academic year, a given Lecture consists of two lecture sections). This was done to account for the different Lecture start times in 7A and 7B during the summer (8:00am and 9:30am, where all summer 7C Lectures were at 8:00am). However, when doing the statistical analyses, it turned out that LecStart in 7A and 7B during the summer is a linear combination of DL11, DL1367, and DL1617 (which represent particular summer DL start times, as described below). In other words, all DLs with one of these three start times were associated with the later summer Lecture time (9:30am) while all DLs with any other summer start times were associated with the earlier summer Lecture time (8:00am). LecStart was therefore omitted from the three statistical analyses involving data from the summer, but conceptually, it is still useful to know that 7A and 7B did have multiple Lecture start times during the summer and that this is accounted for through the (level 2) predictor variables DL11, DL1367, and DL1617.

Student Academics: GPA and Units are both continuous variables that represent, respectively, students' Grade-Point-Averages and the number of units that they got credit for, both specifically at the University prior to the quarter in which the corresponding observation was taken.

Level 2 Predictor Variables

Class Size: The size of different DLs was accounted for and analyzed through a set of seven categorical variables. These are RlySm (under 9 students), Sm (9-14 students), Lit (15-20 students), Med (21-26 students), Stand (27-32 students), Lg (33-38 students), and RlyLg (over 38 students). Stand represents the standard range of sizes for DLs in EPS rooms (more details on different DL rooms below) and is therefore the reference category here. Lit represents a range of DL sizes that fall around the class sizes which are generally considered ideal by the literature and past studies [1]. RlySm (Really Small), Sm (Small), Med (Medium), Lg (Large), and RlyLg (Really Large) are fairly self-explanatory. In EPS DL rooms there are six tables (and therefore, six groups) so it made sense for each of these variables to incorporate six values of possible DL sizes. This is also similar to the ranges used for categorical class size variables in other class size studies [3].

There are two major issues that make it preferable to treat DL size as a categorical variable rather than a continuous one. First, it is possible that the effect of DL size would be non-linear (for instance, if students learn more in classrooms with around 18 students but learn less in classrooms with both greater and fewer numbers of students than this). Secondly, DL sizes are strongly peaked around 30 students (the non-strictly-enforced maximum DL size in EPS rooms) meaning if DL size were treated as a continuous variable, relatively small DL size differences in this range could dominate the statistics when it is not expected that there would be any significant differences between DLs of say, 28 and 31 students (in other words, DLSize is not expected to be continuous in the second sense described in “Levels and Outcome Variable”).

DL Start Time: DL start times are measured in hours based on a 24-hour cycle where minutes are converted to fractional hours given in decimal format and rounded to two decimal places. There are only five possible DL start times during the regular academic year. These are 8:00am (8), 10:30am (10.5), 2:10pm (14.17), 4:40pm (16.67), and 7:10pm (19.17). Because the number of DL start times during the regular academic year is relatively small and one might suspect that the effect of DL start times on Grades would not necessarily be linear (for example, maybe really early and really late start times have a similar effect which is different than the effect of midday start times), or even necessarily a smooth, well-behaved function, it is desirable to treat DL start times as a set of categorical variables rather than a single continuous one. The variables that correspond to the above start times are, respectively, DL8, DL105, DL1417, DL1667, and DL1917.

However, because each DL meets two days per week during the regular academic year there are some instances where a DL will meet at one of the five standard start times on the first day and a different one on the second day. In principle, any combination of two DL start times is possible but the ones that actually occurred in the data are DL start times of 10:30am and 2:10pm, 8:00am and 4:40pm, 10:30am and 4:40pm, and 2:10pm and 4:40pm. In these cases, the two different DL start times (in hours) were averaged to produce that DL's recorded start time with one exception; 8:00am and 4:40pm was recorded as 12.34 instead of 12.33 to distinguish it from the 10:30am and 2:10pm situation. Respectively, the variables corresponding to these situations are DL1233, DL1234, DL1358, and DL1542. The existence of this type of scheduling, which may have effects beyond that of the average start time, is yet another reason to treat DL start times as a set of categorical variables rather than a single continuous one. Each of these situations was treated as its own categorical variable in the analyses, along with categorical variables for each of the standard DL start times (DL8, DL105, DL1417, DL1667, and DL1917 respectively) with DL1417 (2:10pm) serving as the reference category (since during the regular academic year, a large number of observations and DLs are associated with this DL start time, as shown by Tables 9 and 10 in "Analysis Format and Data," and also because theoretically, 2:10pm sections should not have the potential anomalies of sections with a really early or really late start time).

During the summer, the possible DL start times are 9:30am, 11am, 12:10pm, 1:40pm, 2:40pm, 4:10pm, and 5:10pm which corresponded to the variables DL95, DL11, DL1217, DL1367, DL1467, DL1617, and DL1717 respectively while DL1217 served as the reference category (since during the summer, a large number of observations and DLs are associated with this DL start time, as shown by Tables 11 and 12 in "Analysis Format and Data," and also because theoretically, 12:10pm sections should not have the potential anomalies of sections with a really early or a really late start time). During the summer, all four weekly sessions of a given DL always met at the same time.

Classroom: ROS is a dummy variable that has a value of 1 for observations associated with a DL in ROS, a particular building whose rooms have particular layouts, and 0 for those associated with a DL in EPS, which is a different building whose rooms have a different layout than those in ROS. EPS rooms were also specifically designed to accommodate DLs for the courses involved in this study while the overflow room in ROS was modified from its original design to accommodate these DLs. Most DLs for these courses are held in EPS rooms which is why the variable ROS has a value of 0 for observations associated with DLs that take place in EPS.

During the regular academic year, a few 7B DLs met in an EPS room one day per week and a ROS room on the other day. Initially, these situations were accounted for using another dummy variable which was 1 for DLs like this and 0 for those that were not. These DLs were, and are, assigned a ROS value of 0. When doing the statistical analyses, it turned out that this variable is a linear combination of DL1358 and DL1542 (all DLs with either of these start times had one DL per week in a ROS room and the other one in an EPS room and no DLs with any other start times had such an arrangement), so it was omitted from the statistical analyses, but conceptually, it is still useful to know that these types of DLs exist and are accounted for through the predictor variables DL1358 and DL1542.

DL-Mean Predictor Variables: Finally, the DL-means of a few level 1 predictor variables were included as continuous predictor variables at level 2. In particular, Mean_GPA and Mean_Units were included because in an active learning setting, the intention is for students to teach each other, and thus a DL with more overall background knowledge on the part of students (again assuming that grades accurately reflect students' understanding and knowledge of the underlying material) might be expected to positively influence the understanding of new material that individual students in that DL acquire.

Mean_Male was included because, for a variety of social reasons, the number of women in a classroom can also have a significant effect on student learning in that classroom. In this study, the number of women in a given DL will likely be strongly related to Mean_Male (the fraction of students in that DL who identify as male). Male was chosen as the level 1 binary sex variable to take the DL-mean of, rather than Female, because Male directly appears in regressions that include level 1 predictor variables, while Female does not (since Female is the reference category).

Lastly, Mean_LecStart was included for regressions involving level 2 predictor variables and data from the regular academic year (but not those involving data from the summer) because, during the regular academic year, the average Grade in a given DL may be influenced by the fraction of students in that DL who are part of the later lecture section (due to any potential differences between the two lecture sections). During the summer, LecStart is the same for all students in a given DL because they are all part of the same lecture section, and in fact, during the summer all DLs that are part of the same Lecture are part of the same lecture section (which is why during the summer, lecture section and Lecture are equivalent). The effect of lecture section on Grade during the summer is therefore entirely accounted for through the Lecture level of this study's HLM analyses.

Level 3 Predictor Variables

Lecture Size: LecSize is a continuous variable representing the number of students in a given Lecture.

Academic Term: Fall, Winter, and Spring are a set of categorical variables that represent the quarter in which a given DL was held for data from the regular academic year. The reference category here varied depending on the analysis and was Fall for 7A during the regular academic year, Winter for 7B during the regular academic year, and Spring for 7C during the regular academic year. This is because a typical student who enters 7A in the Fall of a given year and takes all three of the courses in this sequence without repeating any of them will take 7B in the Winter and 7C in the Spring, so these reference categories effectively follow this type of typical cohort. Note that all three of the courses in this sequence are offered during all three quarters, as well as both summer sessions, but because many students do not take classes during the summer, while at the same time many do, the cohorts that enter 7A during any other quarter besides the Fall get split up more than the cohorts that enter 7A during a Fall quarter do.

Lecture Instructor: Ins1 – Ins40 are a set of categorical variables representing the instructor who taught (and administered) a given Lecture since most instructors taught multiple Lectures during the period of study. The reference category was different for each regression that included level 3 predictor variables and was defined as the Lecture instructor who had the most actual observations for that regression (“Analysis Format and Summary Data” discusses what is meant by “actual observations”).

Chapter 3: Preliminary Data and Regression Methods

Analysis Format and Summary Data

This study involved six analyses; one for each of the three courses under study (7A, 7B, and 7C) during the regular academic year (Fall, Winter, and Spring quarters) and one for each of these courses during the summer (which includes both Summer Session 1 and Summer Session 2). Each course was analyzed separately for the reasons described in “Level Choices and Predictor Variables” and the summer was analyzed separately from the regular academic year due to the various differences between the two as described in “Site, Sample, and Population.”

As alluded to when discussing course repeats in “Level Choices and Predictor Variables,” observations associated with dropping a course were excluded from the analyses in this study. This is because dropping a course (which initially corresponded to a Grade of 0) is a different outcome than failing it (which also corresponds to a Grade of 0) and even beyond this, dropping a course is a fundamentally distinct outcome from receiving a Grade (so it would not make sense to simply recode these data points as having some Grade other than 0).

The remainder of this section is dedicated to Tables and Figures summarizing the (sample) data used in this study as it relates to the six analyses that were conducted and the predictor variables (with summary data and histograms coming from Version 17 of Stata). Table 1 lists the total number of initial observations for each of the six analyses, the number of observations that were actually part of each analysis (level 1 entities), the number of DLs that were part of each analysis (level 2 entities), and the number of Lectures that were part of each analysis (level 3 entities). The reason that the number of actual observations is different (and less) than the initial total number of observations is because GPA was not recorded in situations where Units was less than 12 (one quarter's worth for a full-time student at the University) and the software used in this study omits observations where at least one variable is missing a value. This means that conclusions drawn from this study will not necessarily apply to first quarter freshmen, transfer students, or graduate students (groups whose members may or may not have taken one of these courses during their first quarter at the University) or to students who do not primarily attend the University (and only took a few courses at the University), because students from these groups are disproportionately excluded from the data used in this study's analyses.

Note that going from 7A to 7B to 7C, during the regular academic year the number of observations and DLs decrease while during the summer they increase. Also note that the number of Lectures is the same for all three courses (25 during the regular academic year and 10 during the summer).

	7A Regular Academic Year	7A Summer	7B Regular Academic Year	7B Summer	7C Regular Academic Year	7C Summer
Total Observations	8317	1009	7401	1199	6136	1453
Actual Observations	7416	970	6840	1178	6043	1441
DLs	293	43	267	52	227	63
Lectures	25	10	25	10	25	10

Table 1: Number of entities at different levels in each of this study's six analyses.

Categorical and Dummy Predictor Variables

The remaining Tables in this section list one of two types of information. First, there are Tables that list the number of actual observations associated with categorical predictor variables, as well as the number of actual observations associated with the "1" value of dummy predictor variables, broken down by level and type of predictor variable. As a reminder for comparison purposes, these Tables include the total number of actual observations for each of the six analyses in this study. Secondly, there are Tables that list the number of level 2 (DL) or level 3 (Lecture) entities that are associated with categorical predictor variables at that level, as well as the number of level 2 or level 3 entities that are associated with the "1" value of dummy predictor variables at that level, broken down by type of predictor variable. As a reminder for comparison purposes, these Tables include, respectively, the total number of level 2 or level 3 entities for each of the six analyses in this study.

In Table 2, note that Repeat does not account for any students who took one or more of these courses prior to the period of study and then retook one or more of them during the period of study, but this number is assumed to be quite small. Also, note how, while some graduate students do take these courses, there are very few of them. Lecture start times are included in Table 2 for all six of this study's analyses, even though during the summer LecStart was omitted from the analyses.

	7A Regular Academic Year	7A Summer	7B Regular Academic Year	7B Summer	7C Regular Academic Year	7C Summer
Actual Observations	7416	970	6840	1178	6043	1441
Later Lecture Start Time	3773	426	3577	571	3235	0
Course Repeats	159	59	582	182	358	151
Graduate Student	8	2	5	1	5	0
Female	4812	644	4423	800	3775	909
Male	2604	326	2417	378	2267	532
Unidentified Binary Sex	0	0	0	0	1	0

Table 2: Numbers of actual observations associated with level 1 categorical and dummy predictor variables (as well as LecStart during the summer), except for those pertaining to race and ethnicity and U.S. citizenship status.

	7A Regular Academic Year	7A Summer	7B Regular Academic Year	7B Summer	7C Regular Academic Year	7C Summer
Actual Observations	7416	970	6840	1178	6043	1441
African-American/Black (AF)	179	40	164	43	152	36
American Indian/Native American (AI)	59	4	58	7	61	9
Chinese-American/Chinese (CH)	1506	200	1392	241	1215	319
East Indian/Pakistani (EI)	544	118	504	133	456	145
Filipino/Filipino-American (FP)	395	64	387	59	313	97
Japanese American/Japanese (JA)	171	15	147	28	116	32
Korean-American/Korean (KO)	217	27	173	49	171	38
Latino/Other Spanish (LA)	279	38	241	45	197	41
Mexican-American/Mexican/Chicano (MX)	832	112	746	139	653	125
Other Asian (OA)	253	40	249	45	219	62
Other (OT)	1	1	5	2	11	2
Pacific Islander. Other (PI)	22	3	22	3	15	8
Vietnamese (VT)	618	87	564	101	495	178
White/Caucasian (WH)	2195	207	2036	265	1834	322
Unidentified Race and Ethnicity (UnE)	145	14	152	18	135	27

Table 3: Numbers of actual observations associated with level 1 categorical predictor variables pertaining to race and ethnicity.

In Table 4, note that almost all of the students who take the courses that were part of this study are U.S. citizens and most of those who are not U.S. citizens are permanent residents.

	7A Regular Academic Year	7A Summer	7B Regular Academic Year	7B Summer	7C Regular Academic Year	7C Summer
Actual Observations	7416	970	6840	1178	6043	1441
U.S. Citizen (Cit)	6637	853	6179	1054	5419	1280
Permanent Resident - Has Green Card (PR)	445	69	406	79	407	104
Visa Holder, Undocumented, or Pending Asylum (NI)	325	47	247	43	207	56
Refugee (RF)	2	0	2	0	3	0
Asylum Granted (PO)	1	0	0	0	0	0
Waiting for Permanent Resident Card -Holds a Valid I-485 Receipt (IM)	2	1	3	2	5	0
Unidentified U.S. Citizenship Status (UnC)	4	0	3	0	2	1

Table 4: Numbers of actual observations associated with level 1 categorical predictor variables pertaining to U.S. citizenship status.

In Tables 5 and 6, note that the overflow DL room was not used all that often during the period of study.

	7A Regular Academic Year	7A Summer	7B Regular Academic Year	7B Summer	7C Regular Academic Year	7C Summer
Actual Observations	7416	970	6840	1178	6043	1441
Overflow DL Room	321	279	19	0	0	290

Table 5: Numbers of actual observations associated with level 2 categorical and dummy predictor variables, except for those pertaining to DL sizes and start times.

	7A Regular Academic Year	7A Summer	7B Regular Academic Year	7B Summer	7C Regular Academic Year	7C Summer
DLs	293	43	267	52	227	63
Overflow DL Room	18	12	1	0	0	14

Table 6: Numbers of DLs (level 2 entities) associated with level 2 categorical and dummy predictor variables, except for those pertaining to DL sizes and start times.

In Tables 7 and 8, note that during the regular academic year, by far the most actual observations come from DLs with a size (number of students) in the Standard range (27 – 32) and by far the most DL sections are of a size that falls into this category. Substantially below the Standard category during the regular academic year are the Medium category (21 – 26) and the Large category (33 – 38). During the summer, the number of actual observations and DL sections in the Medium category are both on par with their respective numbers in the Standard category. Across all six analyses that were part of this study, these categories are followed by the Literature category (15 – 20). Essentially, the Literature category has a decent number of actual observations and DL sections, but still not that many. The Really Small, Small, and Really Large categories (along with the Large category during the summer) have very few actual observations or DL sections associated with them.

	7A Regular Academic Year	**7A Summer**	**7B Regular Academic Year**	**7B Summer**	**7C Regular Academic Year**	**7C Summer**
Actual Observations	7416	970	6840	1178	6043	1441
Really Small	0	0	0	0	8	0
Small	23	0	78	29	112	34
Literature	309	169	200	154	228	272
Medium	676	441	1082	488	1369	454
Standard (in these courses)	5569	327	4706	475	3444	554
Large	802	33	774	32	882	127
Really Large	37	0	0	0	0	0

Table 7: Numbers of actual observations associated with level 2 categorical predictor variables pertaining to DL sizes.

	7A Regular Academic Year	7A Summer	7B Regular Academic Year	7B Summer	7C Regular Academic Year	7C Summer
DLs	293	43	267	52	227	63
Really Small	0	0	0	0	1	0
Small	2	0	7	3	9	3
Literature	18	10	12	9	12	16
Medium	31	20	48	22	58	20
Standard (in these courses)	213	12	176	17	120	20
Large	28	1	24	1	27	4
Really Large	1	0	0	0	0	0

Table 8: Numbers of DLs associated with level 2 categorical predictor variables pertaining to DL sizes.

	7A Regular Academic Year	7B Regular Academic Year	7C Regular Academic Year
Actual Observations	7416	6840	6043
8:00am	1320	1102	766
10:30am	1357	1422	1479
2:10pm	1448	1280	1270
4:40pm	1356	1343	1398
7:10pm	1253	1155	910
10:30am and 2:10pm	658	493	220
8:00am and 4:40pm	24	0	0
10:30am and 4:40pm	0	33	0
2:10pm and 4:40pm	0	12	0

Table 9: Numbers of actual observations associated with level 2 categorical predictor variables pertaining to DL start times during the regular academic year.

	7A Regular Academic Year	7B Regular Academic Year	7C Regular Academic Year
DLs	293	267	227
8:00am	52	46	30
10:30am	52	51	50
2:10pm	58	48	45
4:40pm	54	50	50
7:10pm	51	48	41
10:30am and 2:10pm	25	20	11
8:00am and 4:40pm	1	0	0
10:30am and 4:40pm	0	3	0
2:10pm and 4:40pm	0	1	0

Table 10: Numbers of DLs associated with level 2 categorical predictor variables pertaining to DL start times during the regular academic year.

	7A Summer	7B Summer	7C Summer
Actual Observations	970	1178	1441
9:30am	263	272	542
11:00am	247	261	0
12:10pm	218	213	441
1:40pm	179	186	0
2:40pm	63	122	386
4:10pm	0	124	0
5:10pm	0	0	72

Table 11: Numbers of actual observations associated with level 2 categorical predictor variables pertaining to DL start times during the summer.

	7A Summer	**7B Summer**	**7C Summer**
DLs	43	52	63
9:30am	10	10	20
11:00am	10	10	0
12:10pm	10	10	20
1:40pm	10	10	0
2:40pm	3	7	19
4:10pm	0	5	0
5:10pm	0	0	4

Table 12: Numbers of DLs associated with level 2 categorical predictor variables pertaining to DL start times during the summer.

For level 3 categorical and dummy predictor variable data, note that Lecture instructors are not included in any Table since there are quite a few of them and they were labeled in an anonymous way, meaning it would not provide much useful information to include them.

	7A Regular Academic Year	**7A Summer**	**7B Regular Academic Year**	**7B Summer**	**7C Regular Academic Year**	**7C Summer**
Actual Observations	7416	970	6840	1178	6043	1441
Fall	2591	N/A	1452	N/A	1964	N/A
Winter	3121	N/A	2486	N/A	1310	N/A
Spring	1704	N/A	2902	N/A	2769	N/A

Table 13: Numbers of actual observations associated with level 3 categorical and dummy predictor variables, except for LecStart during the summer and those level 3 categorical predictor variables pertaining to Lecture instructor.

	7A Regular Academic Year	7A Summer	7B Regular Academic Year	7B Summer	7C Regular Academic Year	7C Summer
Lectures	25	10	25	10	25	10
Fall	10	N/A	5	N/A	10	N/A
Winter	10	N/A	10	N/A	5	N/A
Spring	5	N/A	10	N/A	10	N/A

Table 14: Numbers of Lectures associated with level 3 categorical and dummy predictor variables, except for LecStart during the summer and those level 3 categorical predictor variables pertaining to Lecture instructor.

Continuous Predictor Variables

Like with the outcome variable Grade, histograms are the best way to display the distributions of continuous predictor variables in this study, and the Figures in this section do just that (with major deviations discussed in the captions or accompanying text) using actual observations, and for level 2 or level 3 predictor variables, also using level 2 entities (DLs) or level 3 entities (Lectures) respectively.

In Figures 7 and 8, note that GPA and Units are both relatively Normal, but have some skew along with a noticeable ceiling effect for GPA and a noticeable floor effect for Units.

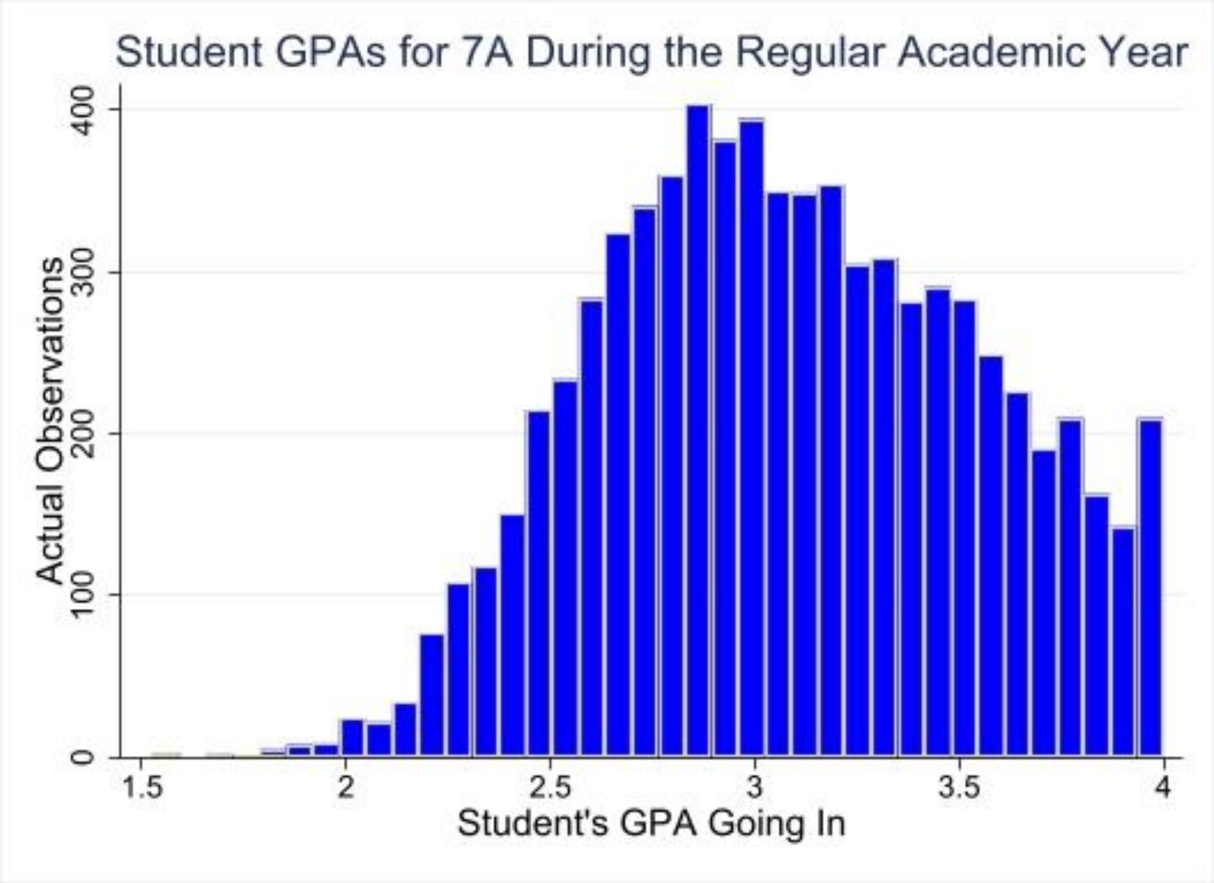

Figure 7: A histogram of GPA during the period of study for 7A during the regular academic year. This distribution is similar to the distribution of GPA in 7A during the summer as well as in 7B and 7C during both the regular academic year and the summer. The main deviations are a flatter peak for 7C during the regular academic year and a slight dip in the peak for 7B during the summer.

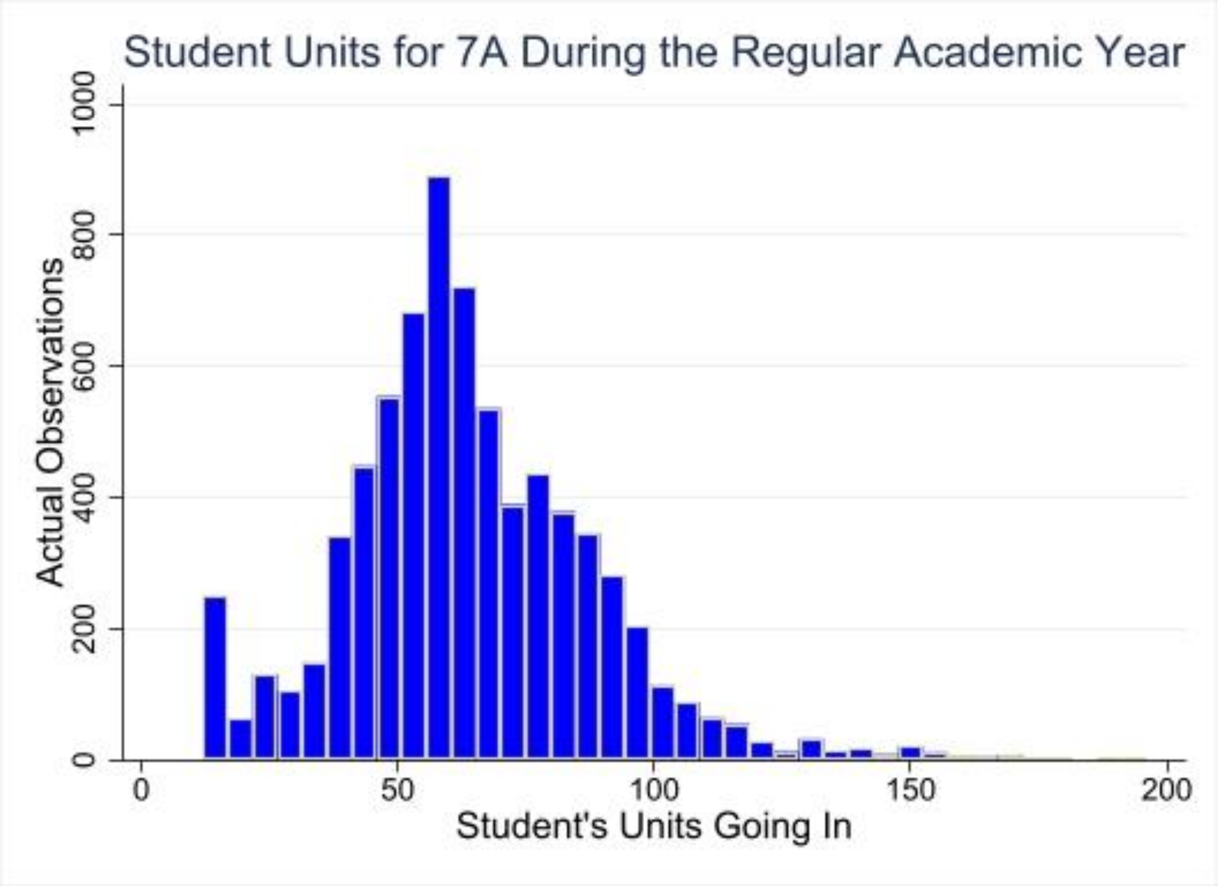

Figure 8: A histogram of Units during the period of study for 7A during the regular academic year. This distribution has a generally similar shape (though with a slightly different peak location) as the distribution of Units in 7A during the summer as well as in 7B and 7C during both the regular academic year and the summer. The main deviations are a less narrow peak for 7B and 7C during the regular academic year and a second peak on the lower end (that is more of a peak than the floor effect shown above) for all three courses during the summer.

In Figures 9, 10, 11, and 12, note that Mean_GPA and Mean_Units follow a distribution that is largely Normal, though for some of the six analyses in this study there is a small dip near the peak.

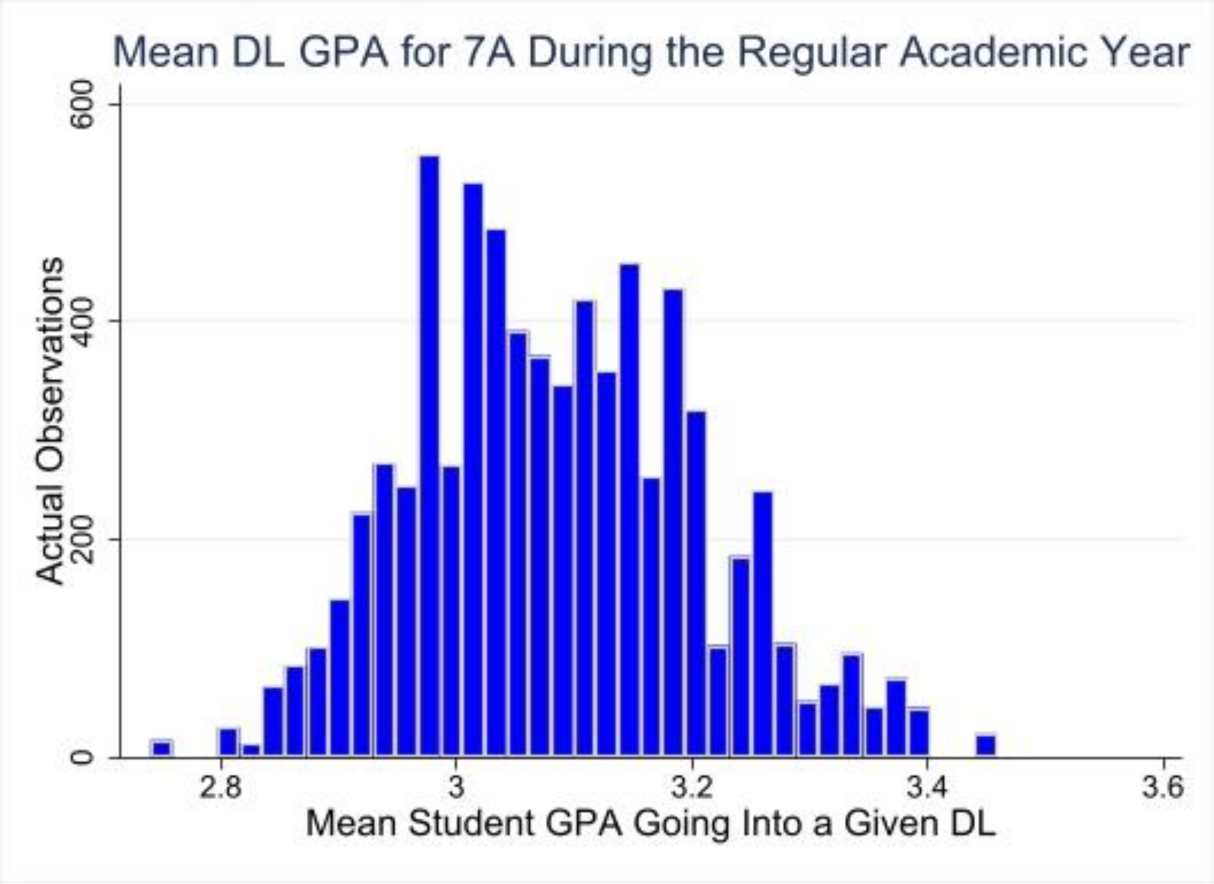

Figure 9: A histogram showing actual observations for Mean_GPA during the period of study for 7A during the regular academic year. This distribution has a similar shape (though with a slightly different peak location) as the distribution of Mean_GPA in 7A during the summer as well as in 7B and 7C during the regular academic year. The main deviation is a standard peak (rather than the dip) for 7B during the regular academic year. 7B during the summer has more of a flat distribution and 7C during the summer has more of an alternating, multiple peaks distribution.

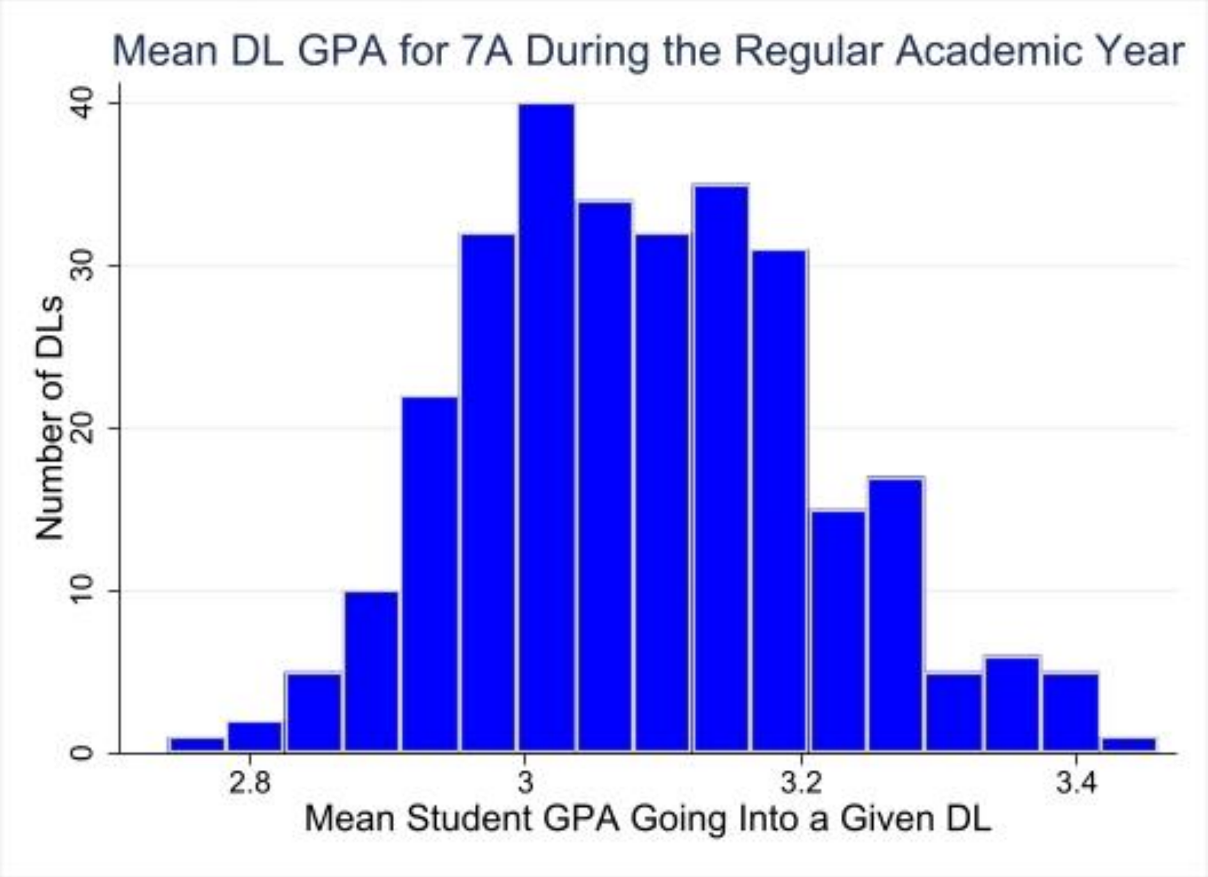

Figure 10: A histogram showing number of DLs for Mean_GPA during the period of study for 7A during the regular academic year. This distribution has a similar shape (though with a slightly different peak location) as the distribution of Mean_GPA in 7B and 7C during the summer, as well as 7C during the regular academic year. The main deviation is a standard peak (rather than the dip) for 7B during the regular academic year and 7A during the summer.

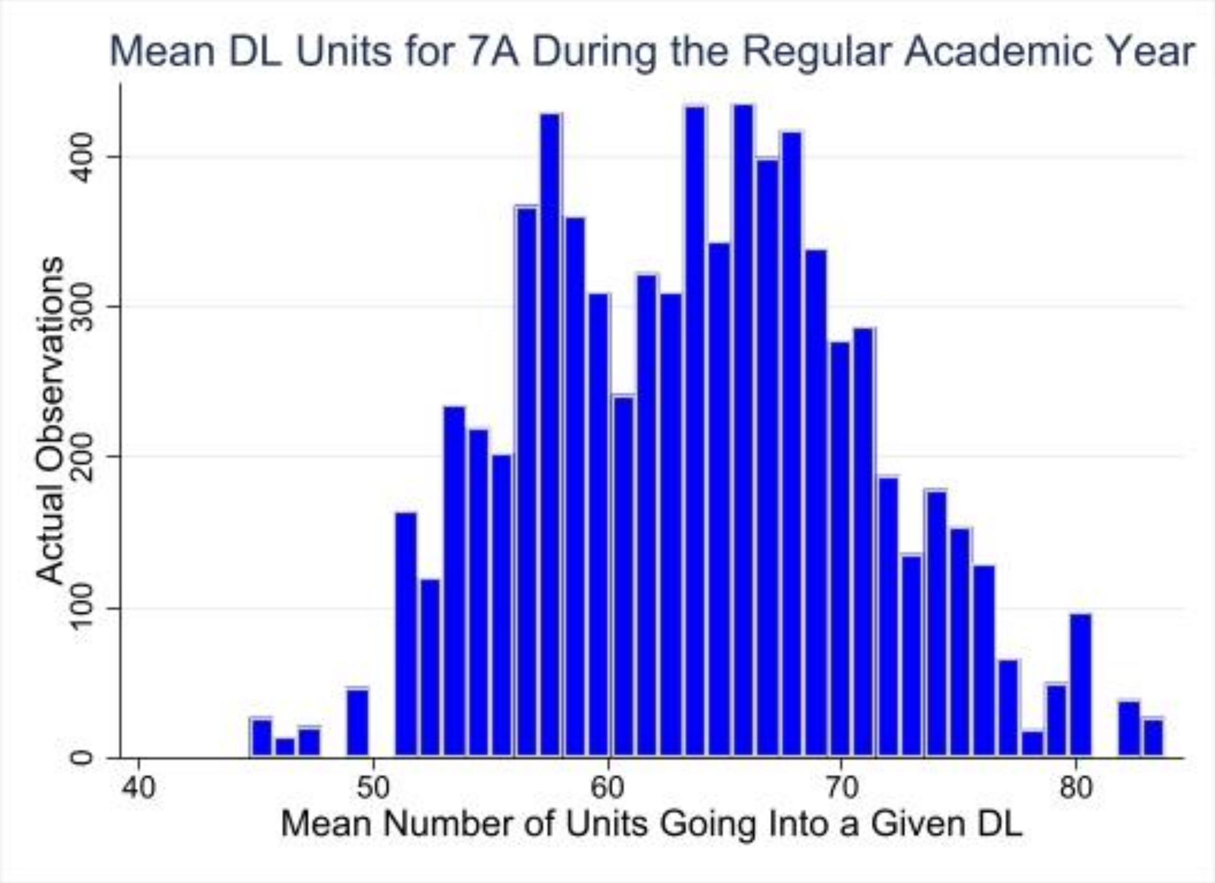

Figure 11: A histogram showing actual observations for Mean_Units during the period of study for 7A during the regular academic year. This distribution has a similar shape (though with a slightly different peak location) as the distribution of Mean_Units in 7A during the summer as well as in 7B and 7C during both the regular academic year and the summer. The main deviations are a standard peak (rather than the dip) for 7A during the summer and 7B during the regular academic year.

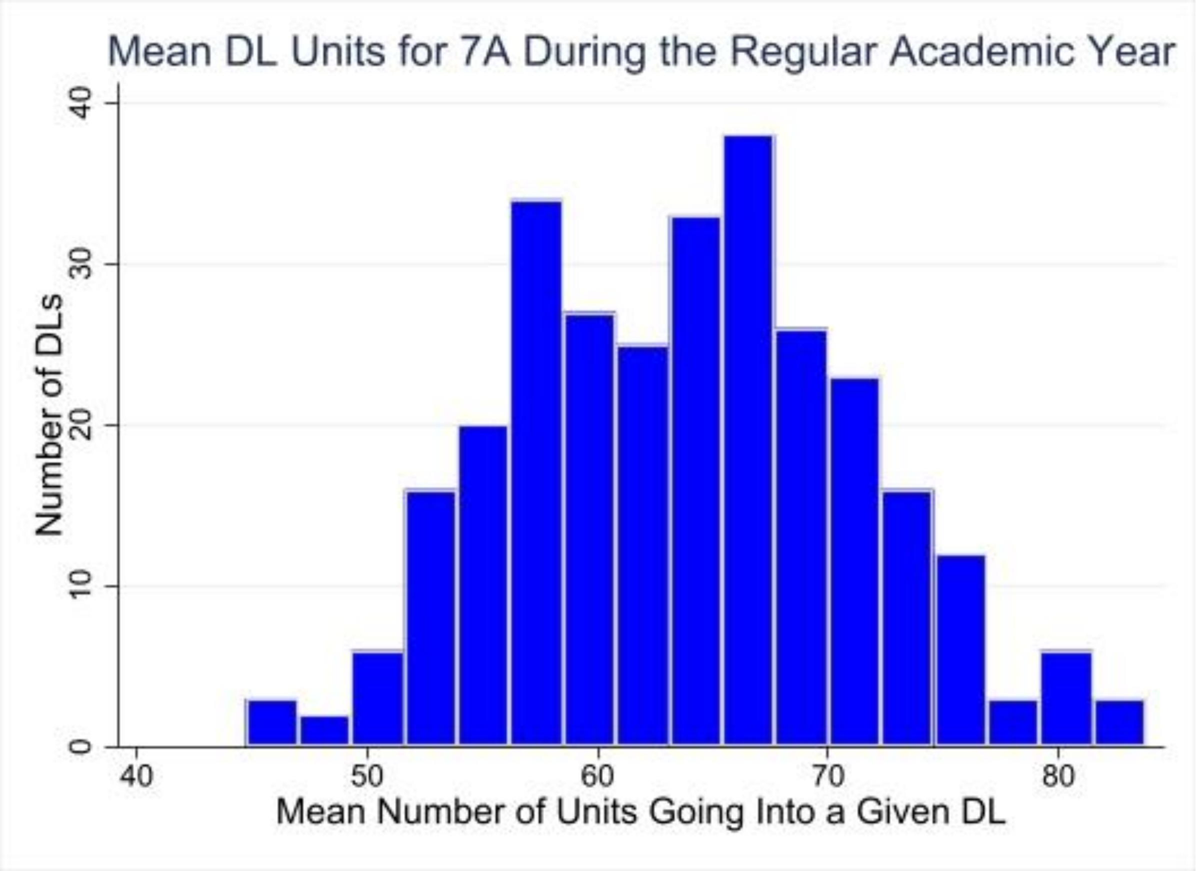

Figure 12: A histogram showing number of DLs for Mean_Units during the period of study for 7A during the regular academic year. This distribution has a similar shape (though with a slightly different peak location) as the distribution of Mean_Units in 7C during the regular academic year, as well as 7B during the summer. The main deviations are a standard peak (rather than the dip) for 7A and 7C during the summer and 7B during the regular academic year.

Note that Mean_Male had a few different distributions across the six analyses in this study, but most of them were mostly Normal. The main difference when it comes to actual observations is in how dips appeared. For 7A during the regular academic year (Figure 13) and 7C during the regular academic year, they were on the sides of the peak creating a fairly standard Normal distribution. For 7B during the regular academic year and 7C during the summer, they were at a location that one might expect to be the peak, which effectively created two peaks. For

7A and 7B during the summer, the distributions had less of a discernible pattern but could be characterized as mostly flat with a few deviations.

When it comes to number of DLs, 7A during the regular academic year (Figure 14) and 7B during the regular academic year followed essentially Normal distributions with small dips at their peaks. 7C during the regular academic year, along with 7B and 7C during the summer, followed an essentially Normal distribution without a dip at the peak. Finally, 7A during the summer followed a highly skewed distribution.

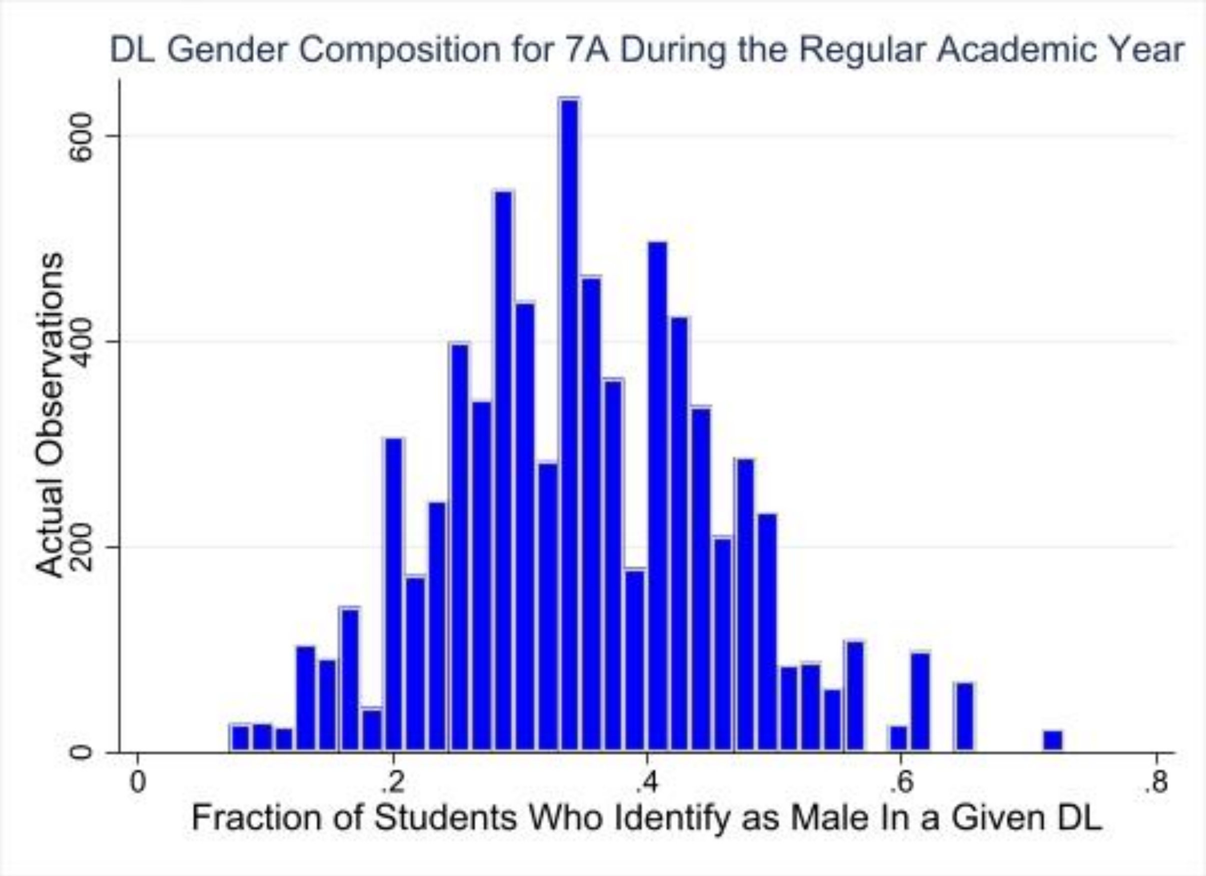

Figure 13: A histogram showing actual observations for Mean_Male during the period of study for 7A during the regular academic year.

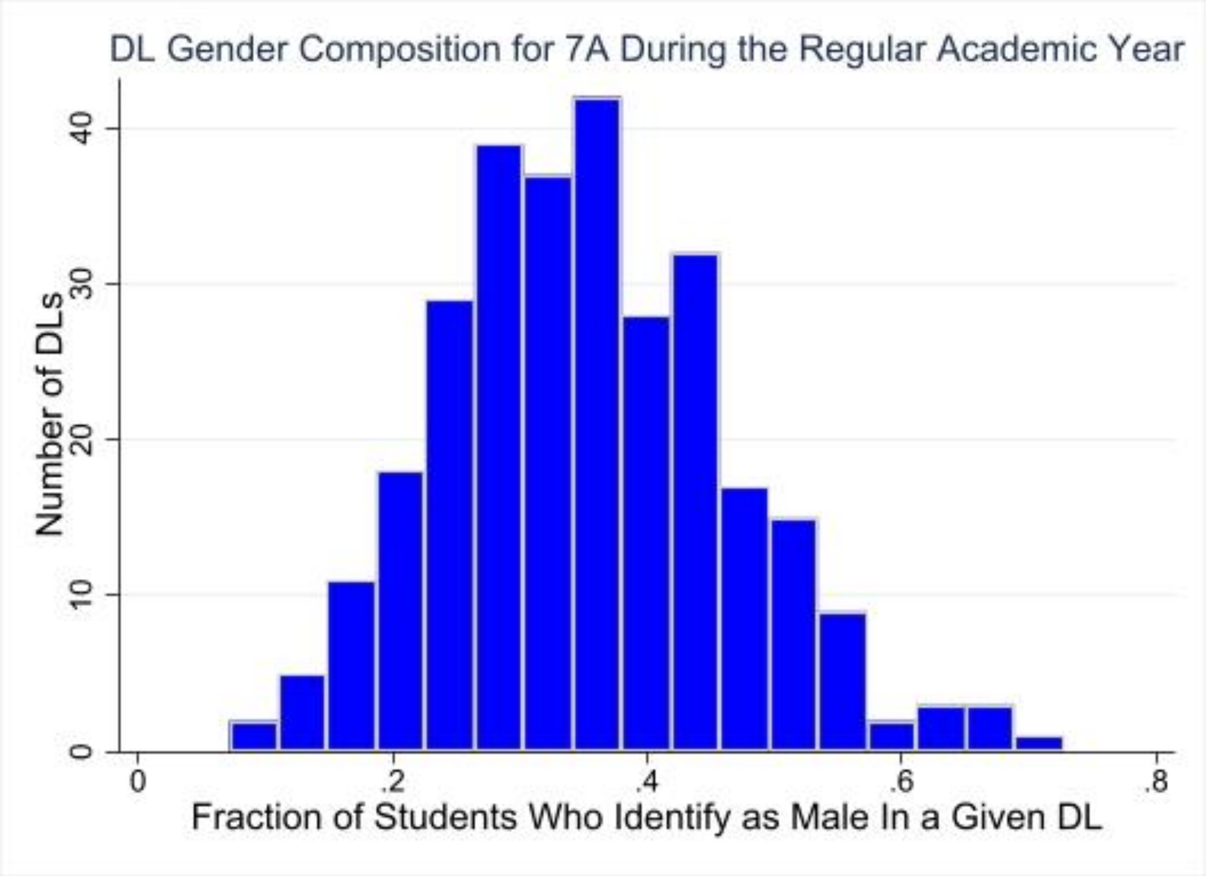

Figure 14: A histogram showing number of DLs for Mean_Male during the period of study for 7A during the regular academic year.

In Figures 15 and 16, note that Mean_LecStart during the regular academic year is sharply peaked around 0.5 with half of the students in a given DL coming from the earlier lecture section (which as a reminder, is part of a given Lecture) and half coming from the later lecture section, though there are some small dips, and this distribution gets a bit wider going from 7A to 7B to 7C. The situation during the summer is very different for the reasons mentioned in the Mean_LecStart part of “Level Choices and Predictor Variables.”

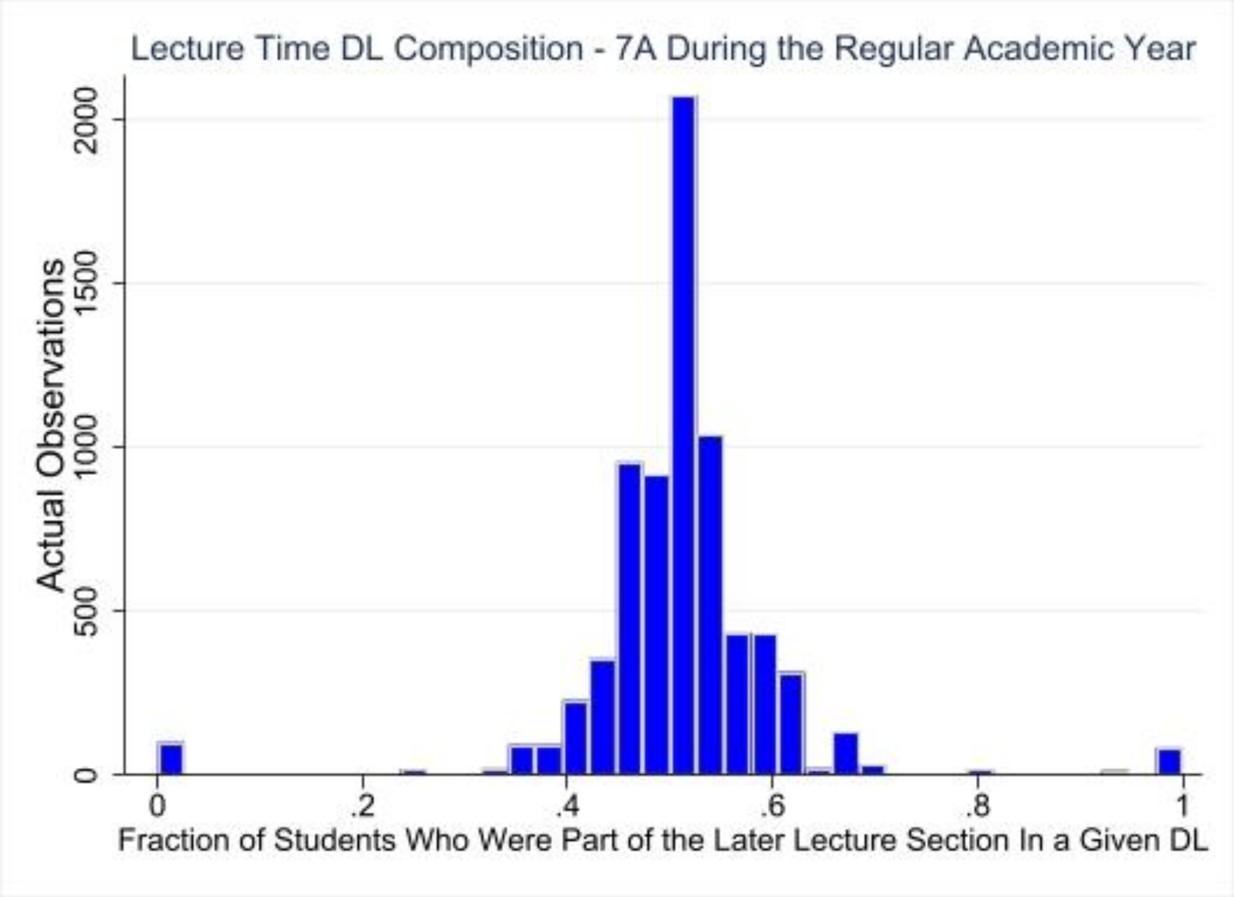

Figure 15: A histogram showing actual observations for Mean_LecStart during the period of study for 7A during the regular academic year. This distribution is similar to the distribution for Mean_LecStart in 7B and 7C during the regular academic year.

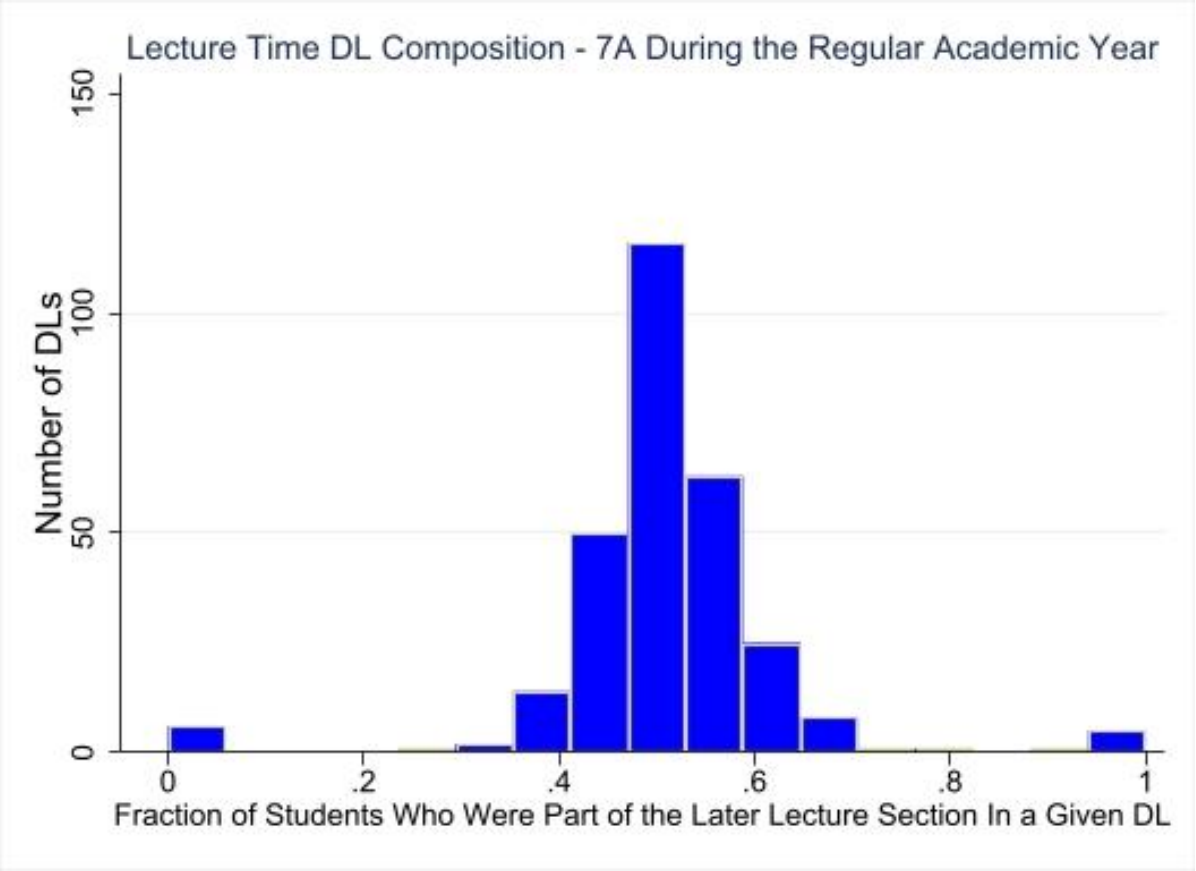

Figure 16: A histogram showing number of DLs for Mean_LecStart during the period of study for 7A during the regular academic year. This distribution is similar to the distribution for Mean_LecStart in 7B and 7C during the regular academic year.

LecSize does not really follow a set pattern between the six analyses, but when it comes to actual observations, each distribution can broadly be described as consisting of one or two flat distributions with a few deviations. 7A during the regular academic year (Figure 17) and 7B during the regular academic year essentially have two superimposed flat distributions (a lower one and a higher one), along with a few deviations. 7A during the summer, 7C during the regular academic year, and 7C during the summer essentially have one flat distribution, along with a few deviations. 7B during the summer has an almost perfectly flat distribution.

When it comes to number of Lectures, 7A during the regular academic year (Figure 18) follows a loosely Normal distribution while 7C during the regular academic year essentially has two superimposed flat distributions (a lower one and a higher one), along with a few deviations. 7A during the summer (Figure 19), 7B during the summer, and 7C during the summer, as well as 7B during the regular academic year, follow highly skewed distributions.

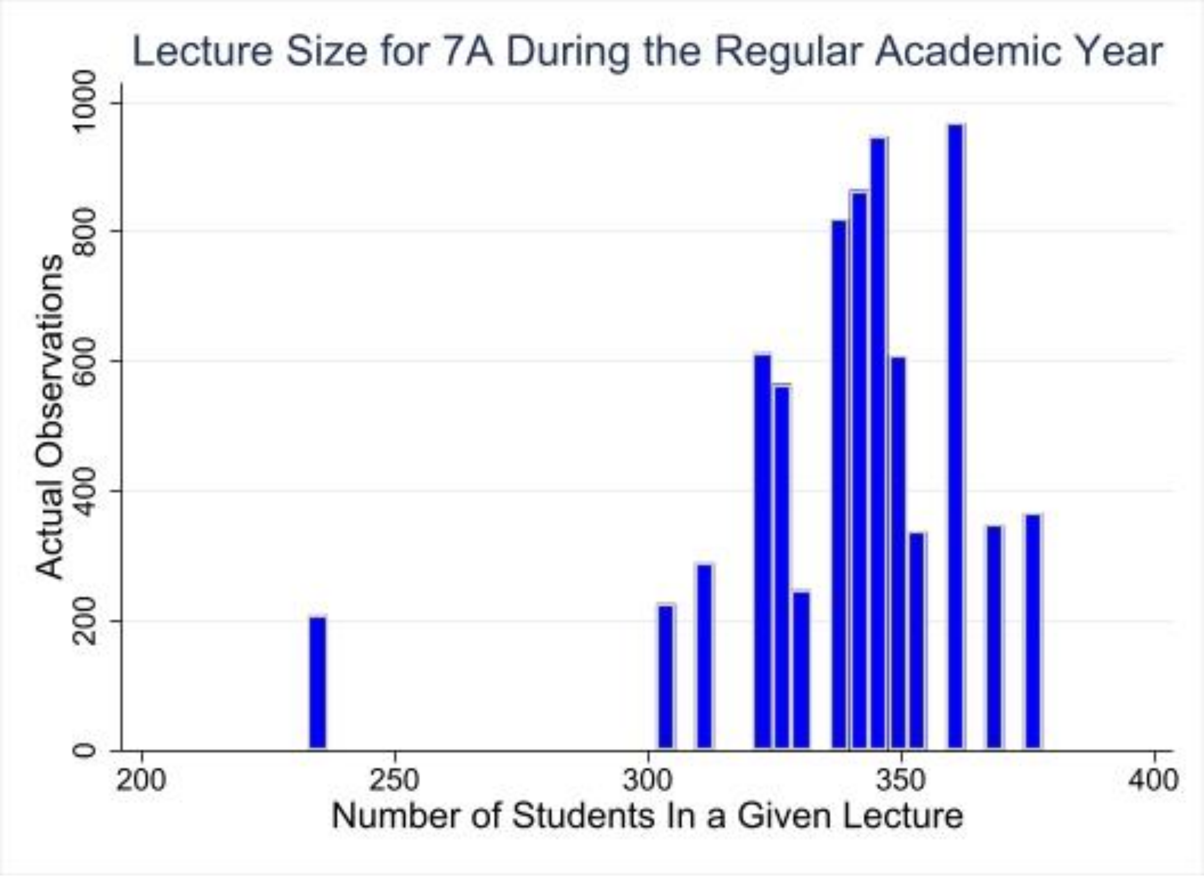

Figure 17: A histogram showing actual observations for LecSize during the period of study for 7A during the regular academic year.

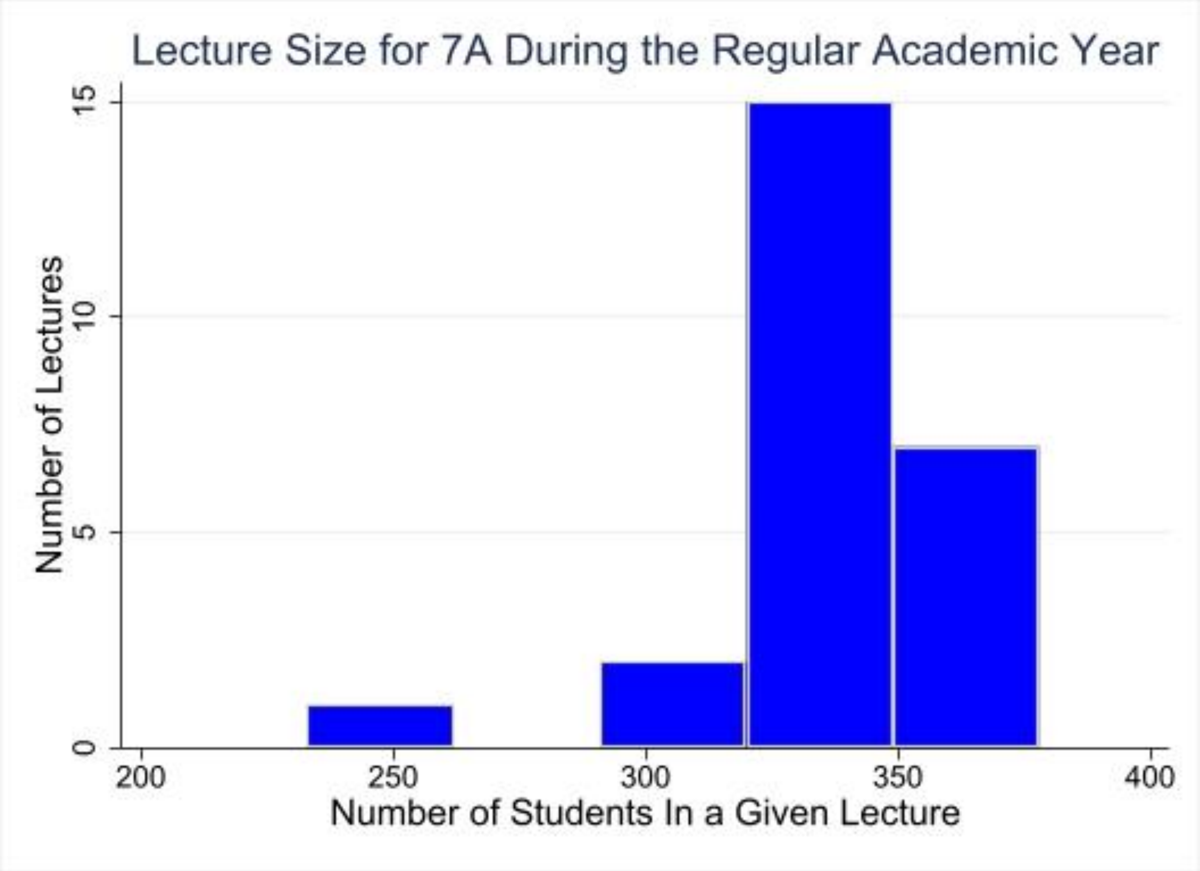

Figure 18: A histogram showing number of Lectures for LecSize during the period of study for 7A during the regular academic year.

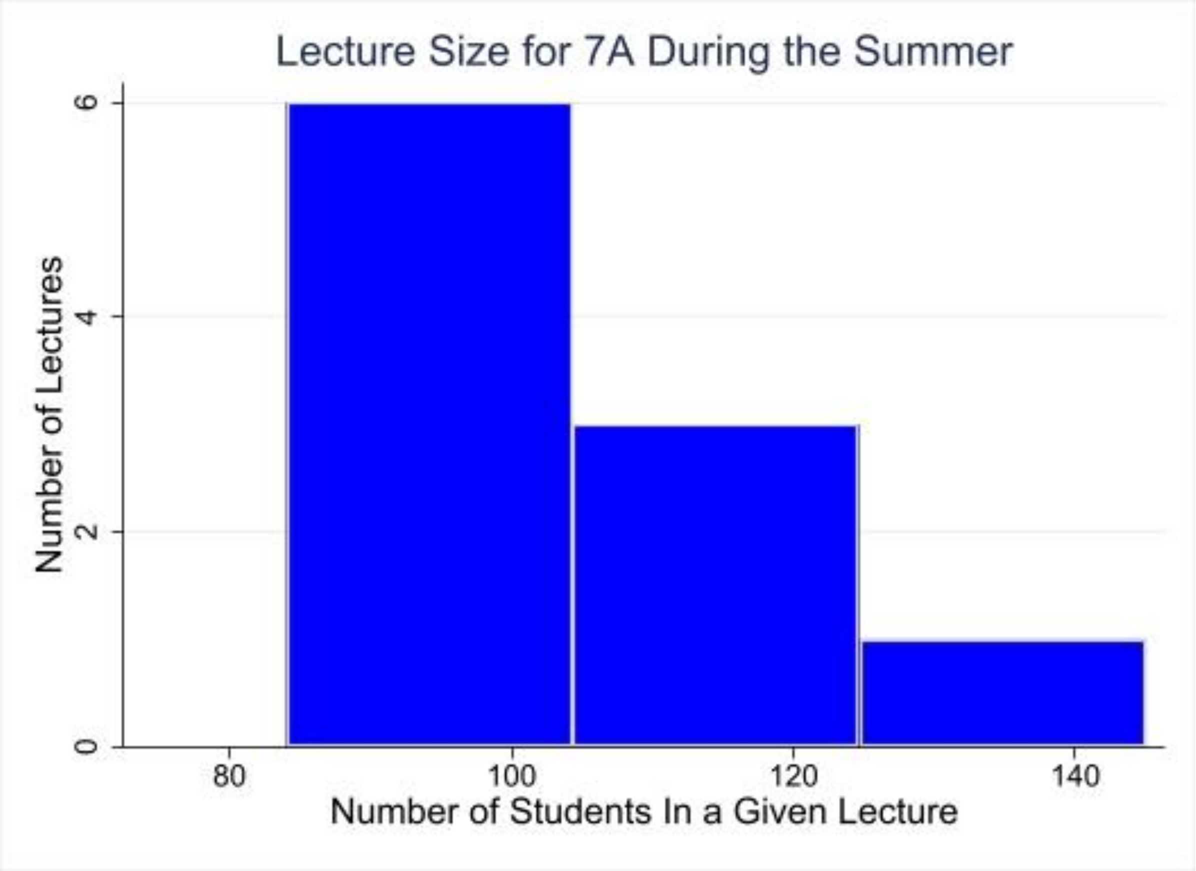

Figure 19: A histogram showing number of Lectures for LecSize during the period of study for 7A during the summer.

Analysis Procedure and Methods

Each of the six analyses in this study was conducted in four stages (regressions) that were done on actual observations (as defined in "Analysis Format and Summary Data") using Version 17 of Stata. Each of these stages involved random intercepts, meaning the intercepts (the value of Grade when all predictor variables have a value of 0) were allowed to vary between different DLs and Lectures, but none of these regressions included random slopes, meaning the slopes on predictor variables were not allowed to vary between different DLs or Lectures (check out "Statistical Techniques and Random Effects" and Appendix A for more details about random intercepts and random slopes in Multilevel Modeling, as well as HLM specifically). Slopes were fixed in this study because there were no strong theoretical arguments for slopes to vary between either DLs or Lectures nor were there any prominent potential interactions between different predictor variables. Check out Appendix C for more details about the HLM equations that were used at each stage of this study.

As is the case with most HLM analyses, the goal of the first stage of each of the six analyses in this study was to determine the amount of total initial variance in the outcome variable (Grade in this study), as well as how much initial variance (as both an amount and as a fraction of the total initial variance) exists at each level. In other words, a major goal of this first stage is to determine how much initial variance in the outcome variable (Grade) lies between observations within a given level 2 entity (a DL in this study), between different level 2 entities (DLs) within a given level 3 entity (a Lecture in this study), and between different level 3 entities (Lectures). These variances are determined through a Null Model, which is a model that does not include any predictor variables, so the first stage of each of the six analyses in this study involved specifying and fitting a Null Model.

The goal of the second stage was to control for the effects of level 1 predictor variables and to determine what impact they have on Grade (as well as what impact they have on the variance in Grade at different levels) when no level 2 or level 3 predictor variables are included. To do this, an Individual Model was specified and fit which included all level 1 predictor variables, but no predictor variables from any other levels.

Similarly, the goal of the second stage was to control for the effects of level 2 predictor variables and to determine what impact they have on Grade (as well as what impact they have on the variance in Grade at different levels) when no level 3 predictor variables are included.

Therefore, at this stage a DL Model was specified and fit which included all level 1 and level 2 predictor variables, but no level 3 predictor variables.

The goal of the final stage was twofold. The first goal was similar to the goals of the previous two stages; to control for the effects of level 3 predictor variables and determine what impact they have on Grade (as well as what impact they have on the variance in Grade at different levels). However, the other main goal of this final stage was to finalize a determination of the effect that each predictor variable (all of them from all three levels) had on Grade (by examining their corresponding slopes and slope uncertainties). Theoretically, the presence of predictor variables from a given level should not affect the slopes or slope uncertainties that are associated with predictor variables from other levels, nor should the presence of predictor variables from a given level affect the variance in Grade at other levels, but in practice, the presence of predictor variables from a given level does sometimes affect aspects of the regression that are associated with other levels. Thus, at this stage, a Final Model was specified and fit which included all predictor variables (from levels 1, 2, and 3). Of the various slopes and slope uncertainties that were computed in this Final Model, the most important ones for this study (because they are the ones that directly relate to this study's main research question) are those associated with the set of DL size categorical predictor variables and, though to a lesser degree, those associated with the continuous Lecture size predictor variable.

Chapter 4: Results, Discussion, and Conclusions

Regression Results and Variance Discussion

This section is devoted to listing and describing results (variances in Grade and slopes on predictor variables) from the regressions that were part of this study. It also includes interpretations around the meaning of Grade variances, intraclass correlation coefficients (ICCs), and variance changes in these results, but does not include interpretations around the meaning of slopes, which will be examined in “Slope Discussion and Overall Conclusions.” While a regression was conducted for each of the four stages of each of the six analyses in this study, the slopes on predictor variables in the Individual and DL models are similar to those in the Final Model and the main reason Individual and DL Model regressions were run was to determine the resulting changes in Grade variance at each level (where theoretically, including a given level’s corresponding predictor variables should reduce the variance in Grade at that level, but not at other levels). Because of this, slopes from the Individual and DL Models are not included here, but slopes from the Final Model are, along with Grade variances and related quantities (ICCs and changes in variance) at each level of each model. In all of the Tables in this section, the standard error for a given quantity is in parentheses either below or to the right of that quantity. These standard errors were either taken directly from the regression software (Stata version 17) or were calculated using error propagation techniques combined with standard errors taken directly from the regression software.

Null Model

Table 15 shows the initial variance in Grades at each level for each of the six analyses in this study (based on the Null Model, where no predictor variables were included in the corresponding regressions), along with initial DL and Lecture ICCs for each of these analyses (respectively, the fraction of total initial variance between different DLs within a given Lecture and the fraction of total initial variance between different Lectures). Check out the end of Appendix A for how ICCs are calculated. It is generally expected that the vast majority of initial total variance in an HLM analysis will occur between level 1 entities within a given level 2 entity, and that is indeed the case here where for all six of the analyses in this study, the vast majority of initial total variance in Grade is between observations within a given DL (which is itself within a given Lecture).

However, for all six of these analyses, the proportion of initial total variance in Grade that is between DLs within a given Lecture is quite low (well below 0.05, the informal bar that is often used when examining Null Model ICCs in HLM analyses). Furthermore, the standard errors in the DL ICCs are rather high, with only one that is less than half of the corresponding ICC value (7C during the summer). This is somewhat surprising when most of the class-time and class-based learning in the courses that were part of this study occurs during DL. On the other hand, the proportion of initial total variance in Grade that exists between different Lectures is rather large (well above 0.05) for five out of six analyses (the exception being 7B during the regular academic year). This is not too surprising considering Grades in these courses are largely determined by quizzes and exams, which are typically written by Lecture instructors, as well as administered during lecture sections by Lecture instructors (as discussed in “Site, Sample, and Population”), and a Lecture’s instructor also determines their Lecture’s overall grading scheme (albeit with relatively little variation between different instructors, as mentioned in “Levels and Outcome Variable”).

	7A Regular Academic Year	7A Summer	7B Regular Academic Year	7B Summer	7C Regular Academic Year	7C Summer
Initial Student/Observation Variance	0.692 (0.012)	0.622 (0.029)	0.896 (0.016)	0.584 (0.025)	0.740 (0.014)	0.806 (0.031)
Initial DL Variance	0.0037 (0.0027)	0.0129 (0.0099)	0.0064 (0.0038)	0.0143 (0.0092)	0.0009 (0.0029)	0.031 (0.013)
Initial Lecture Variance	0.054 (0.016)	0.098 (0.048)	0.0232 (0.0076)	0.072 (0.036)	0.136 (0.039)	0.082 (0.041)
ICC2 (DL)	0.0049 (0.0036)	0.018 (0.013)	0.0069 (0.0041)	0.021 (0.013)	0.0010 (0.0033)	0.034 (0.014)
ICC3 (Lecture)	0.072 (0.020)	0.133 (0.058)	0.0251 (0.0081)	0.108 (0.048)	0.155 (0.038)	0.089 (0.041)

Table 15: Initial variances and ICCs for each level of each of the six analyses in this study.

Individual Model

Table 16 shows the variance at each level for each of the six analyses in this study after implementing the Individual Model (when level 1, student based, predictor variables were included in the corresponding regressions), along with percent changes in these variances relative to the Null Model, which were calculated as:

$$\%\text{ Change in Variance From Null Model} = \frac{\text{New Variance} - \text{Initial Variance}}{\text{Initial Variance}} * 100$$

Note that for all six analyses, the level 1 variance decreased substantially (between 36% and 49%). Variances at the DL and Lecture levels changed as well, which should not have been the case, but they did not change in any systematic way and sometimes this happens in these regressions, especially with small variances. Furthermore, with the exception of the DL variance for 7C during the summer, the standard error in the percent changes for all of the DL and Lecture variances were quite large compared to the corresponding values of these changes. Another possible factor here for the three analyses involving data from the regular academic year is that, for logistical reasons, LecStart is being treated as a level 1 predictor variable in these analyses when it would make more theoretical sense as a level 3 predictor variable.

	7A Regular Academic Year	**7A Summer**	**7B Regular Academic Year**	**7B Summer**	**7C Regular Academic Year**	**7C Summer**
Individual Model Student/Observation Variance	0.3932 (0.0066)	0.350 (0.016)	0.4574 (0.0080)	0.372 (0.016)	0.4026 (0.0075)	0.486 (0.019)
Individual Model DL Variance	0.0030 (0.0016)	0.0196 (0.0089)	0.0036 (0.0020)	0.0117 (0.0066)	0.0018 (0.0017)	0.0042 (0.0057)
Individual Model Lecture Variance	0.045 (0.013)	0.119 (0.057)	0.0232 (0.0071)	0.076 (0.037)	0.154 (0.044)	0.085 (0.040)
% Change in Student/Observation Variance From Null Model	-43.2 (1.3)	-43.8 (3.7)	-48.9 (1.3)	-36.3 (3.8)	-45.6 (1.4)	-39.8 (3.3)
% Change in DL Variance From Null Model	-20. (73)	50. (140)	-45 (45)	-18 (70.)	120 (750)	-86 (19)
% Change in Lecture Variance From Null Model	-17 (35)	22 (84)	0. (45)	6 (73)	13 (46)	5 (72)

Table 16: Individual Model variances and percent changes in variances from their initial values for each level of each of the six analyses in this study.

DL Model

Table 17 shows the variance at each level for each of the six analyses in this study after implementing the DL Model (when predictor variables from levels 1 and 2, meaning both student and DL based predictor variables, were included in the corresponding regressions), along with percent changes in these variances relative to the Null Model (where these percent changes were calculated in the same way as above). Note that for all six analyses, the level 1 variances remained largely the same as they were in the Individual Model. Most level 2 variances, however, decreased substantially relative to the Null Model (between 45% and 100%), with the exception being 7C during the regular academic year (where the level 2 variance increased relative to the Null Model). However, the standard error in the percent changes of these DL level variances is quite high compared to their corresponding values for four of the six analyses (the exceptions being 7B and 7C during the summer). Here it is also important to mention that 7C during the regular academic year had a particularly low amount of initial DL variance (even less than the other five analyses), which makes the increase in its variance likely the result of statistical noise (further evidence of this can be found in the extremely high standard error for the percent change in this variance). Level 3 variances changed as well, which should not have been the case, but they did not change in any systematic way and statistical fluctuations are once again the likely culprit, especially given the high standard errors in these percent changes compared to their corresponding values.

	7A Regular Academic Year	7A Summer	7B Regular Academic Year	7B Summer	7C Regular Academic Year	7C Summer
DL Model Student/Observation Variance	0.3931 (0.0066)	0.349 (0.016)	0.4573 (0.0080)	0.372 (0.016)	0.4026 (0.0075)	0.481 (0.018)
DL Model DL Variance	0.0018 (0.0015)	0.0070 (0.0054)	0.0026 (0.0019)	0.0020 (0.0044)	0.0010 (0.0016)	0.0000000000 (3.6E-9)
DL Model Lecture Variance	0.046 (0.014)	0.082 (0.041)	0.0237 (0.0073)	0.078 (0.038)	0.153 (0.044)	0.097 (0.045)
% Change in Student/Observation Variance From Null Model	-43.2 (1.3)	-43.9 (3.7)	-48.9 (1.3)	-36.3 (3.8)	-45.6 (1.4)	-40.4 (3.2)
% Change in DL Variance From Null Model	-50. (54)	-45 (59)	-59 (38)	-86 (32)	20 (440)	-100.000000 (1.1E-5)
% Change in Lecture Variance From Null Model	-14 (36)	-16 (59)	2 (46)	8 (75)	13 (46)	18 (82)

Table 17: DL Model variances and percent changes in variances from their initial values for each level of each of the six analyses in this study.

Final Model

Table 18 shows the variance at each level for each of the six analyses in this study after implementing the Final Model (when all predictor variables from all three levels, meaning those corresponding to students, DLs, and Lectures, were included in the corresponding regressions), along with percent changes in these variances relative to the Null Model (where these percent changes were calculated in the same way as above). Note that for all six analyses, the level 1 variances remained largely the same as they were in the Individual and DL Models. Four of the DL level variances also remained largely the same as they were in the DL Model while two of them (7A and 7B during the summer) decreased substantially (though likely as a result of statistical fluctuations given how low these variances are and how high the corresponding standard errors are). All Lecture level variances decreased substantially relative to the Null Model (between 84% and 100%), as well as relative to the Individual and DL Models.

	7A Regular Academic Year	7A Summer	7B Regular Academic Year	7B Summer	7C Regular Academic Year	7C Summer
Final Model Student/ Observation Variance	0.3932 (0.0066)	0.349 (0.016)	0.4574 (0.0080)	0.371 (0.015)	0.4026 (0.0075)	0.477 (0.018)
Final Model DL Variance	0.0019 (0.0015)	0.0056 (0.0044)	0.0025 (0.0019)	0.000000000 (2.7E-8)	0.0010 (0.0016)	0.000000000 (1.1E-8)
Final Model Lecture Variance	0.00130 (0.00085)	3E-11 (4.6E-10)	0.0035 (0.0016)	2E-16 (2.1E-15)	0.0114 (0.0038)	0.00000000 (3.8E-7)
% Change in Student/ Observation Variance From Null Model	-43.2 (1.3)	-44.0 (3.7)	-48.9 (1.3)	-36.4 (3.7)	-45.6 (1.4)	-40.8 (3.2)
% Change in DL Variance From Null Model	-50. (54)	-57 (48)	-61 (37)	-100.00000 (0.00019)	20. (450)	-100.000000 (3.5E-5)
% Change in Lecture Variance From Null Model	-97.6 (1.7)	-99.99999997 (4.7E-7)	-84.9 (8.4)	-99.9999999999998 (2.9E-12)	-91.6 (3.7)	-100.00000 (0.00046)

Table 18: Final Model variances and percent changes in variances from their initial values for each level of each of the six analyses in this study.

The remainder of this section is devoted to listing and describing slopes from the Final Model for each of the six analyses in this study, broken down by level and types of predictor variables. An "N/A" slope in the following Tables means that the slope does not exist because the corresponding predictor variable was not part of that regression, either for fundamental reasons (like how neither Mean_LecStart nor Winter were part of any regressions involving data from the summer) or because all of the actual observations in a particular regression happened to have the same value for that variable (i.e. 0 since all of these situations involved dummy or categorical predictor variables). In these Tables, slopes that are statistically significant at the 95%, 99%, or 99.9% confidence levels are indicated, respectively, with an asterisk (*), two asterisks (**), or three asterisks (***) to the right of them. Any reference to a slope being statistically significant means that it is statistically significant at the 95% confidence level or higher. Since the interpretations of many of these slopes are quite similar, a few example interpretations will be discussed, but not all of the slopes will be mentioned.

Predictor Variable	7A Regular Academic Year	7A Summer	7B Regular Academic Year	7B Summer	7C Regular Academic Year	7C Summer
GPA	1.123*** (0.017)	1.016*** (0.043)	1.160*** (0.019)	0.884*** (0.042)	1.094*** (0.019)	1.088*** (0.043)
Units	-0.00066* (0.00032)	-0.00010 (0.00064)	-0.00168*** (0.00031)	-0.00036 (0.00056)	-0.00075** (0.00026)	-0.00017 (0.00054)
Repeat	-0.186*** (0.051)	0.050 (0.083)	-0.814*** (0.031)	-0.231*** (0.053)	-0.751*** (0.036)	-0.521*** (0.062)
Grad	0.05 (0.22)	-0.02 (0.43)	0.31 (0.30)	-0.01 (0.62)	-0.51 (0.29)	N/A
LecStart	0.022 (0.015)	N/A	0.011 (0.017)	N/A	0.042* (0.017)	N/A
Male	0.192*** (0.016)	0.174*** (0.042)	0.239*** (0.018)	0.184*** (0.040)	0.203*** (0.017)	0.145*** (0.039)
Female	Reference	Reference	Reference	Reference	Reference	Reference
UnS	N/A	N/A	N/A	N/A	2.06* (0.90)	N/A

Table 19: Slopes on level 1 predictor variables, except for those pertaining to race and ethnicity and U.S. citizenship status, in the Final Model.

For continuous predictor variables, the corresponding slope is the average amount by which the outcome variable (Grade in this study) changes as a result of a one unit change in that predictor variable, after accounting for the effects of all other predictor variables. For instance, in Table 19 the slope on GPA (a continuous level 1 predictor variable) for 7A during the regular academic year is 1.123 (and is statistically significant). This means that in 7A during the regular academic year, an increase of one GPA point (on a 4.00 scale, which is what GPA is measured in), and thus, an increase of one letter grade (for instance, from a C+ to a B+), for a given student is associated with an average increase of 1.123 Grade points (on a 4.00 scale, which is what Grade is measured in) for that student after accounting for all other predictor variables.

For sets of categorical predictor variables, the slope on a given one of these is the average amount by which the outcome variable is different for observations which are affiliated with that category compared to observations that are affiliated with the reference category, after accounting for the effects of all other predictor variables. For instance, in Table 19 the slope on Male (which is part of a set of level 1 categorical predictor variables) for 7A during the regular academic year is 0.192 (and is statistically significant). This means that in 7A during the regular academic year, students who identified as Male received a Grade that was on average 0.192 points higher than the Grade received by students who identified as Female (the reference category) after accounting for all other predictor variables.

For dummy predictor variables, the corresponding slope is the average amount by which the outcome variable is different for observations which are assigned a value of "1" compared to those which are assigned a value of "0," after accounting for the effects of all other predictor variables. For instance, in Table 19 the slope on Repeat (a dummy level 1 predictor variable) for 7A during the regular academic year is -0.186 (and is statistically significant). This means that in 7A during the regular academic year, students who had previously taken 7A for a letter grade at least once during the period of study (i.e. those with a Repeat value of 1) received a Grade that was on average 0.186 points lower than the Grade received by students who were taking 7A for a letter grade for the first time during the period of study (i.e. those with a Repeat value of 0) after accounting for all other predictor variables.

Predictor Variable	7A Regular Academic Year	7A Summer	7B Regular Academic Year	7B Summer	7C Regular Academic Year	7C Summer
AF	-0.254*** (0.049)	-0.28** (0.10)	-0.293*** (0.055)	-0.09 (0.10)	-0.220*** (0.054)	-0.22 (0.12)
AI	-0.008 (0.083)	-0.24 (0.30)	0.011 (0.090)	0.02 (0.24)	0.105 (0.083)	-0.41 (0.24)
CH	0.006 (0.022)	-0.010 (0.062)	0.072** (0.024)	0.091 (0.056)	-0.046 (0.024)	-0.083 (0.056)
EI	-0.109*** (0.030)	-0.102 (0.070)	-0.104** (0.034)	-0.053 (0.066)	-0.181*** (0.034)	-0.028 (0.070)
FP	-0.064 (0.035)	-0.148 (0.086)	-0.095* (0.038)	-0.047 (0.089)	-0.090* (0.039)	-0.019 (0.081)
JA	0.065 (0.050)	-0.13 (0.16)	0.000 (0.058)	0.19 (0.12)	-0.020 (0.061)	0.30* (0.13)
KO	-0.185*** (0.046)	-0.07 (0.12)	-0.134* (0.055)	-0.007 (0.098)	-0.193*** (0.052)	0.00 (0.12)
LA	-0.278*** (0.040)	-0.29** (0.11)	-0.146** (0.046)	-0.01 (0.10)	-0.163*** (0.048)	-0.07 (0.12)
MX	-0.222*** (0.026)	-0.304*** (0.071)	-0.170*** (0.030)	-0.139* (0.066)	-0.213*** (0.029)	-0.128 (0.074)
OA	-0.096* (0.042)	-0.12 (0.11)	-0.082 (0.046)	0.02 (0.10)	-0.148*** (0.046)	-0.140 (0.098)
OT	0.36 (0.63)	-0.18 (0.60)	0.07 (0.30)	-0.13 (0.44)	-0.14 (0.19)	0.00 (0.49)
PI	-0.16 (0.14)	-0.20 (0.35)	-0.17 (0.15)	0.21 (0.36)	-0.25 (0.17)	0.08 (0.25)
VT	-0.100*** (0.029)	0.020 (0.078)	-0.025 (0.033)	-0.004 (0.073)	-0.159*** (0.032)	-0.018 (0.066)
WH	Reference	Reference	Reference	Reference	Reference	Reference
UnE	-0.047 (0.055)	0.19 (0.16)	0.016 (0.058)	-0.06 (0.15)	-0.082 (0.057)	0.22 (0.14)

Table 20: Slopes on race and ethnicity predictor variables in the Final Model.

Predictor Variable	7A Regular Academic Year	7A Summer	7B Regular Academic Year	7B Summer	7C Regular Academic Year	7C Summer
Cit	Reference	Reference	Reference	Reference	Reference	Reference
PR	0.051 (0.032)	-0.039 (0.077)	0.060 (0.036)	0.008 (0.074)	0.061 (0.034)	0.044 (0.073)
NI	0.159*** (0.038)	-0.004 (0.097)	0.226*** (0.046)	0.27** (0.10)	0.164*** (0.047)	0.293** (0.099)
RF	-0.55 (0.45)	N/A	-0.52 (0.48)	N/A	-0.34 (0.37)	N/A
PO	-0.06 (0.63)	N/A	N/A	N/A	N/A	N/A
IM	0.00 (0.45)	-0.25 (0.60)	0.03 (0.39)	0.53 (0.44)	0.42 (0.29)	N/A
UnC	0.02 (0.32)	N/A	0.40 (0.40)	N/A	-0.56 (0.64)	-0.27 (0.71)

Table 21: Slopes on U.S. citizenship status predictor variables in the Final Model.

Predictor Variable	7A Regular Academic Year	7A Summer	7B Regular Academic Year	7B Summer	7C Regular Academic Year	7C Summer
ROS	-0.005 (0.064)	-0.056 (0.072)	-0.20 (0.21)	N/A	N/A	-0.059 (0.063)
Mean_GPA	-0.121 (0.078)	-0.61** (0.21)	0.001 (0.096)	1.50*** (0.30)	0.00 (0.12)	0.28 (0.19)
Mean_Units	-0.0009 (0.0015)	-0.0044 (0.0038)	-0.0011 (0.0014)	0.0032 (0.0030)	-0.0005 (0.0012)	-0.0043 (0.0034)
Mean_Male	-0.079 (0.086)	0.10 (0.30)	-0.104 (0.091)	0.64** (0.24)	0.05 (0.10)	-0.10 (0.24)
Mean_LecStart	0.056 (0.083)	N/A	0.02 (0.15)	N/A	-0.23 (0.14)	N/A

Table 22: Slopes on level 2 predictor variables, except for those pertaining to DL sizes and start times, in the Final Model.

In Table 22, note that the slope on Mean_GPA (a continuous level 2 predictor variable) for 7A during the summer is -0.61 (and is statistically significant). This means that an increase of one Mean_GPA point (on a 4.00 scale, which is what Mean_GPA is measured in) within a given 7A DL during the regular academic year is associated with an average decrease of 0.61 Grade points for all students in that DL after accounting for all other predictor variables.

Predictor Variable	7A Regular Academic Year	7A Summer	7B Regular Academic Year	7B Summer	7C Regular Academic Year	7C Summer
RlySm	N/A	N/A	N/A	N/A	0.08 (0.24)	N/A
Sm	0.04 (0.14)	N/A	0.03 (0.13)	0.16 (0.14)	-0.002 (0.071)	0.41** (0.15)
Lit	0.051 (0.066)	-0.16 (0.12)	-0.041 (0.060)	-0.058 (0.076)	-0.003 (0.050)	-0.136 (0.081)
Med	0.008 (0.033)	-0.193* (0.081)	-0.007 (0.028)	-0.115* (0.057)	-0.009 (0.024)	-0.026 (0.067)
Stand	Reference	Reference	Reference	Reference	Reference	Reference
Lg	-0.064* (0.027)	0.04 (0.16)	-0.047 (0.033)	0.14 (0.15)	0.009 (0.029)	-0.018 (0.082)
RlyLg	-0.14 (0.12)	N/A	N/A	N/A	N/A	N/A

Table 23: Slopes on level 2 predictor variables pertaining to DL sizes in the Final Model.

In Table 23, note that the slope on Med for 7A during the summer is -0.193 (and is statistically significant). This means that in 7A during the summer, all students who were part of a given DL that had a size in the Medium range/category received a Grade that was an average of 0.193 points lower than the Grade received by students who were part of a DL in the Standard range/category (the reference category) after accounting for all other predictor variables.

Predictor Variable	7A Regular Academic Year	7B Regular Academic Year	7C Regular Academic Year
DL8	0.040 (0.026)	0.003 (0.032)	0.028 (0.034)
DL105	0.002 (0.026)	0.021 (0.029)	0.021 (0.026)
DL1417	Reference	Reference	Reference
DL1667	0.017 (0.026)	0.013 (0.029)	0.037 (0.026)
DL1917	0.042 (0.029)	0.053 (0.033)	0.070* (0.033)
DL1233	-0.011 (0.032)	0.002 (0.040)	0.038 (0.054)
DL1234	0.16 (0.14)	N/A	N/A
DL1358	N/A	0.02 (0.18)	N/A
DL1542	N/A	0.32 (0.25)	N/A

Table 24: Slopes on level 2 predictor variables pertaining to DL start times during the regular academic year in the Final Model.

Predictor Variable	7A Summer	7B Summer	7C Summer
DL95	-0.123 (0.077)	-0.151* (0.077)	0.016 (0.061)
DL11	-1.56*** (0.27)	-0.88*** (0.13)	N/A
DL1217	Reference	Reference	Reference
DL1367	-1.48*** (0.30)	-0.83*** (0.14)	N/A
DL1467	-0.08 (0.11)	-0.076 (0.082)	0.124* (0.055)
DL1617	N/A	-0.93*** (0.14)	N/A
DL1717	N/A	N/A	0.04 (0.10)

Table 25: Slopes on level 2 predictor variables pertaining to DL start times during the summer in the Final Model.

Predictor Variable	7A Regular Academic Year	7A Summer	7B Regular Academic Year	7B Summer	7C Regular Academic Year	7C Summer
LecSize	-0.00077 (0.00053)	-0.0196** (0.0076)	0.00088 (0.00079)	-0.0780*** (0.0087)	0.0006 (0.0029)	0.0028 (0.0026)
Fall	Reference	N/A	0.18 (0.12)	N/A	0.02 (0.29)	N/A
Winter	-0.282*** (0.082)	N/A	Reference	N/A	-0.31* (0.12)	N/A
Spring	-0.447*** (0.083)	N/A	-0.092 (0.080)	N/A	Reference	N/A

Table 26: Slopes on level 3 predictor variables, except for those pertaining to Lecture instructor, in the Final Model.

In Table 26, note that the slope on LecSize (a continuous level 3 predictor variable) for 7A during the summer is -0.0196 (and is statistically significant). This means that an increase of one LecSize unit (i.e. one student in a Lecture) for a given 7A Lecture during the summer is associated with an average decrease of 0.0196 Grade points for all students in that Lecture after accounting for all other predictor variables.

Also note that the slope on Spring for 7A during the regular academic year is -0.447 (and is statistically significant). This means that in 7A during the regular academic year, all students who were part of a given Lecture that occurred in a Spring quarter received a Grade that was an average of 0.447 points lower than the Grade received by students who were part of a Lecture that occurred in a Fall quarter (the reference category) after accounting for all other predictor variables.

Predictor Variable	7A Regular Academic Year	7A Summer	7B Regular Academic Year	7B Summer	7C Regular Academic Year	7C Summer
Ins1	N/A	N/A	-0.29* (0.14)	N/A	N/A	N/A
Ins2	N/A	N/A	-0.31* (0.14)	N/A	N/A	N/A
Ins3	-0.541*** (0.069)	N/A	0.383*** (0.095)	N/A	N/A	N/A
Ins4	-0.058 (0.082)	N/A	N/A	N/A	N/A	N/A
Ins5	N/A	N/A	N/A	N/A	N/A	-0.278*** (0.082)
Ins6	N/A	N/A	N/A	N/A	0.68*** (0.17)	N/A
Ins7	N/A	N/A	0.43* (0.18)	Reference	N/A	N/A
Ins8	0.049 (0.074)	N/A	N/A	N/A	-0.39 (0.27)	N/A
Ins9	N/A	N/A	N/A	N/A	-0.12 (0.16)	N/A
Ins10	N/A	N/A	N/A	N/A	0.57** (0.19)	N/A
Ins11	0.001 (0.072)	Reference	0.18 (0.14)	0.76*** (0.10)	0.58** (0.20)	N/A
Ins12	N/A	N/A	-0.18 (0.14)	N/A	N/A	N/A
Ins13	-0.115 (0.074)	N/A	N/A	N/A	N/A	N/A
Ins14	N/A	N/A	N/A	N/A	N/A	Reference
Ins15	0.006 (0.074)	N/A	N/A	N/A	N/A	N/A
Ins16	N/A	N/A	Reference	N/A	Reference	N/A
Ins17	-0.260** (0.083)	0.21 (0.19)	N/A	N/A	N/A	-0.140 (0.071)
Ins18	N/A	N/A	0.29* (0.12)	N/A	N/A	0.613*** (0.098)
Ins19	Reference	N/A	N/A	N/A	N/A	N/A
Ins20	N/A	N/A	0.310*** (0.090)	N/A	N/A	N/A
Ins21	N/A	N/A	N/A	0.640*** (0.085)	N/A	N/A
Ins22	0.230** (0.082)	1.14*** (0.22)	N/A	N/A	N/A	N/A

Ins23	N/A	N/A	N/A	N/A	N/A	-0.540*** (0.082)
Ins24	-0.102 (0.075)	-0.01 (0.11)	0.192* (0.088)	1.06*** (0.15)	N/A	N/A
Ins25	-0.074 (0.064)	N/A	0.08 (0.10)	N/A	-0.18 (0.22)	N/A
Ins26	0.485*** (0.076)	N/A	N/A	N/A	N/A	N/A
Ins27	N/A	N/A	N/A	N/A	-0.14 (0.18)	N/A
Ins28	-0.080 (0.073)	N/A	N/A	N/A	N/A	-0.286** (0.097)
Ins29	N/A	N/A	N/A	1.90*** (0.17)	N/A	N/A
Ins30	N/A	-0.54*** (0.10)	N/A	N/A	N/A	N/A
Ins31	-0.053 (0.076)	N/A	N/A	N/A	N/A	N/A
Ins32	-0.045 (0.072)	0.31 (0.31)	N/A	N/A	N/A	N/A
Ins33	N/A	N/A	N/A	N/A	-0.35 (0.20)	-0.36** (0.12)
Ins34	N/A	N/A	N/A	N/A	0.90*** (0.15)	N/A
Ins35	N/A	N/A	N/A	N/A	N/A	-0.31 (0.21)
Ins36	-0.412*** (0.081)	0.89*** (0.25)	N/A	1.50*** (0.19)	N/A	N/A
Ins37	-0.077 (0.061)	N/A	N/A	N/A	N/A	N/A
Ins38	N/A	N/A	N/A	N/A	0.54* (0.24)	N/A
Ins39	N/A	N/A	N/A	0.71*** (0.10)	N/A	N/A
Ins40	N/A	N/A	N/A	N/A	-0.04 (0.19)	N/A

Table 27: Slopes on level 3 predictor variables pertaining to Lecture instructor in the Final Model.

In Table 27, note that the reference categories were different for each of this study's six analyses (with one overlap). They were Ins19 for 7A during the regular academic year, Ins11 for 7A during the summer, Ins16 for 7B during the regular academic year, Ins7 for 7B during the summer, Ins16 for 7C during the regular academic year, and Ins14 for 7C during the summer.

Also note that there are a lot of N/As (basically, a lot of the Lecture instructors for the three courses that were part of this study only taught Lecture for one or two of these courses and/or were only the Lecture instructor for such courses during either the regular academic year or the summer, but not both).

Slope Discussion and Overall Conclusions

Given the large number of slopes in the Final Model, it would not be productive to discuss all of them. Therefore, the discussion (and conclusions) here will be focused on patterns, along with slopes that are related to the main research question around the effect of class size on student understanding of physics concepts (which Grade was used as a proxy for). A major aspect of potential patterns that will be considered here is statistical significance (at the 95% confidence level or higher), though this will not be the only metric that is used since, while it is a commonly accepted and fairly effective standard (for finding what predictor variables are relevant and meaningful), it is still rather arbitrary and far from perfect (especially since there are far more than 20 predictor variables in this study, as well as for some of the reasons discussed below).

Another major aspect of different slopes' potential patterns and importance that will be considered here is the magnitude (absolute value) of these slopes, sometimes described as an effect size, because if the magnitude of a given predictor variable's corresponding slope is too small, then that predictor variable does not have much of an effect on the outcome variable, even if its corresponding slope is statistically significant. For categorical and dummy predictor variables (where slopes represent the difference in Grade between different groups after accounting for all other predictor variables), in this study a small slope magnitude refers to a slope with a magnitude that is less than 0.11 (i.e. 1/3 of 0.33, where 0.33 is the lowest gap between two consecutive Grades). For continuous predictor variables, what constitutes a small slope magnitude depends on the typical range of values for that slope's corresponding predictor variable. There are different ways of deciding on, and accounting for, this typical range, but since this study does not involve many continuous predictor variables, their corresponding slopes will be discussed on a case-by-case basis as needed (though the general idea will always be about comparing Grade changes due to a given predictor variable to 0.11).

When it comes to the slopes on level 1 predictor variables, it is unsurprising that University GPA consistently has a large positive association with Grade (considering a student's University GPA is essentially just a weighted average of their past grades at the University). There was also a consistently negative association between Units and Grade, which was statistically significant for all three courses in this study during the regular academic year, though not during the summer. Furthermore, while the magnitudes of these slopes may seem small, they are actually substantial given the range of Units in the data. For instance, in 7A during the regular academic year, the minimum value of Units is 12 while the maximum value is 196. This means that, after accounting for the effects of all other predictor variables, a student who went into 7A during the regular academic year with the maximum number of Units received a Grade that was on average 0.12 points (on a 4.00 scale) lower than that of a student who went in with the minimum number of Units. It is a bit surprising that students who have more Units going into the courses that were part of this study during the regular academic year receive lower Grades in them despite these students presumably having more experience with University courses and also more background knowledge in relevant areas. However, it is true that the design of these courses is different than most University courses and the material covered in them, while related to material covered in other courses that bioscience majors at the University take, is not that strongly tied to such material. In fact, part of the reason for the large range of Units that students have going into these courses is because these courses are not pre-requisites for other courses that bioscience majors at the University are required to take (except each other since 7A is a pre-requisite for 7B and 7B is a pre-requisite for 7C). It is also the case that bioscience majors sometimes put off these courses and perhaps those who go in with more Units have, on average, more aversion to such courses than those who go in with fewer Units. Finally, students who go into these courses with more Units are presumably more likely to be taking upper-level courses simultaneously while those who go in with fewer Units are presumably more likely to be taking lower-level courses simultaneously, and perhaps this affects the amount of time and energy that different students can devote to these courses.

Repeat had a fairly sizable and consistent (for five out of six analyses), as well as statistically significant, negative association with Grade, meaning students who previously took one of the three courses in this study for a letter grade and then retook that course got lower Grades than those who did not retake that course. This is somewhat surprising because these students had already gone through the course material (and format), which in some ways gives them an advantage when encountering the material and format again. However, it is also quite possible that whatever factor(s) caused them to get a low enough Grade the first time they took a course that they decided to retake it affected their Grade during subsequent takes as well. During the regular academic year, the later Lecture start time is consistently associated with higher Grades, but this effect is only statistically significant for one course (7C) and the magnitudes of the corresponding slopes are quite small, so the practical impact this has on Grades is minimal even if it is a real effect (as opposed to the result of statistical fluctuations).

Given previous research and general knowledge about U.S. society, the reader will probably be unsurprised to learn that being Female (as compared to Male) or from certain marginalized and underrepresented racial and ethnic groups (as compared to White) gives a student with such identities systematic disadvantages, as evidenced in this study by consistently positive (and statistically significant) slopes corresponding to the Male predictor variable (where Female was the reference category) and consistently negative slopes corresponding to predictor variables representing certain marginalized and underrepresented racial or ethnic groups. Some of these slopes are statistically significant and some of them are not, but the pattern is there and in some cases, the lack of statistical significance is likely at least partially a result of disproportionately low numbers of students from these groups taking the courses that were part of this study, which is itself often related to systematic, and systemic, disadvantages. The same is true of Refugees (RF), who had systematically lower Grades than U.S. citizens (as with all slopes, after accounting for all other predictor variables) even though the resulting slopes were not statistically significant. Given the low number of students in the RF category, these negative slopes could be statistical fluctuations, but they could also constitute a real pattern and be the result of structural factors. Furthermore, since the slope corresponding to a given predictor variable is determined in the context of controlling for all other predictor variables, students who identify as Female, part of certain marginalized and underrepresented racial or ethnic groups, or as a refugee are experiencing disadvantages which are not accounted for through the other predictor variables in this study. Therefore, these students are likely experiencing additional disadvantages as well, just ones which are obscured by this study's other predictor variables. For instance, since a student's University GPA is essentially just a weighted average of their past grades at the University and these students have systematically lower grades in the courses that were part of this study, it is likely that they also have disproportionately low University GPAs. If this is the case, then the related issues are obscured by controlling for GPA.

On the other hand, the slopes corresponding to students who were visa holders, undocumented, or pending asylum (NI) were fairly consistently positive and statistically significant (the exception being 7A during the summer), meaning these students consistently received higher grades than U.S. citizens. Note, though, that since the NI predictor variable represents a rather broad and diverse category, these slopes do not necessarily indicate much about any particular group that was part of this category. Finally, the slope corresponding to Unidentified Binary Sex for 7C during the regular academic year was statistically significant and rather large. However, also note that 7C during the regular academic year was the only analysis (out of the six that were conducted as part of this study) which had any actual observations in this category, and further note that it only had one, so this slope probably does not have much meaning (beyond an individual meaning for the single observation that it corresponds to) and should serve as yet another cautionary example against automatically deferring to statistical significance without additional context. It is important, though, to recognize that the lack of non-binary gender options in the data that was used for this study is itself a problem and likely contributes to this study's inability to gauge the effect that identifying outside the gender binary has on Grades.

When it comes to level 2 predictor variables (excluding those around DL size for now), the overflow DL room had a consistently negative association with Grade (for the four out of six regressions that included DLs which were held in this room). However, none of the corresponding slopes were statistically significant and all but one of them had small magnitudes anyway (the exception being 7B during the regular academic year). None of the slopes corresponding to the four continuous variables that are DL means of level 1 predictor variables seemed to follow any consistent pattern (even though a few of them were statistically significant).

During the regular academic year, all but one of the DL start times had a positive association with Grade compared to the 2:10pm reference category (the exception being the situation in 7A where one of that DL's two weekly sessions meets at 10:30am while the other one meets at 2:10pm). This is somewhat surprising when 2:10pm (i.e. early afternoon) seems like a time when students would be the most attentive, but this may not be the case after considering the rest of their schedules. Furthermore, only one of the corresponding slopes was statistically significant (the exception being 7:10pm for 7C) and all but two of them were rather small (the exceptions being the situation where one DL session per week meets at 8:00am and the other one meets at 4:40pm for 7A and the situation where one DL session per week meets at 2:10pm and the other one meets at 4:40pm for 7B). During the summer, the association between DL start times and Grade (compared to the 12:10pm reference category) was largely mixed (despite quite a few of the corresponding slopes being statistically significant). However, the 11:00am, 1:40pm, and 4:10pm start times all had a consistently negative, statistically significant, and fairly large association with Grade. Furthermore, recall that these DL start times are all affiliated with the later Lecture start time during the summer, meaning it seems like during the summer, students in the later Lecture start time got significantly lower Grades than students in the earlier Lecture start time. It is unknown why this is though. It could have to do with the DL times themselves, it could have to do with the Lecture time itself, it could have to do with one or more other factors that are correlated with these times, or it could have to do with a combination thereof.

When it comes to level 3 predictor variables (excluding the one around Lecture size for now), during the regular academic year, Grades were consistently lower during the Winter and Spring quarters than during the Fall quarter. More specifically, for 7A where Fall was the reference category, the slopes on the Winter and Spring predictor variables were both negative and statistically significant. For 7B, where Winter was the reference category, the slope on the Fall predictor variable was positive while the slope on the Spring predictor variable was negative (though neither were statistically significant) and for 7C, where Spring was the reference category, the slope on the Fall predictor variable was positive (though not statistically significant) while the slope on the Winter predictor variable was negative (and was statistically significant), One possible reason for this is that most students come into Fall quarter energized and rested after the summer (at least in terms of academic work), whereas Winter quarter is known around the University to be the worst quarter (and follows a winter break that typically only lasts about three weeks) and Spring quarter follows a spring break of only one week.

There were also statistically significant and rather large associations between Grade and many of the Lecture instructors across the six analyses that were part of this study, which makes sense given the fundamental and inherent relationship between Lecture instructors and Grades, as discussed in “Site, Sample, and Population” (among other places). There is also evidence to suggest that the grading scale used in a given Lecture plays a role when determining Grades, and that even when the same Lecture instructor uses different grading scales, the outcomes are often quite different [14]. However, during the period that this study’s data was drawn from, most Lecture instructors were using the same grading scale. Furthermore, if grading scale is a factor here, its effect is controlled for through the Lecture level of an HLM framework along with the Lecture instructor categorical predictor variables. Therefore, since any such effects are not related to this study’s primary research question, it is not necessary to distinguish them from the effects of the Lecture level or Lecture instructors for the purposes of this study. It is interesting and relevant to note, though, that the traditional percentage grading scale seems to cause more variation in Grades between Lectures than the alternative 4.5 point grading scale does (where the 4.5 point grading scale corresponds directly to standard letter grades while the percentage grading scale needs to be converted to standard letter grades) [14].

Class Size

The association between Grade and both DL and Lecture size was rather mixed and did not follow any clear patterns. Compared to the standard DL size reference category (27-32 students), across the six analyses that were part of this study, there were both positive and negative slopes corresponding to the DL size categorical predictor variables Sm (9-14 students), Lit (15-20 students), Med (21-26 students), and Lg (33-38 students). Note that it was not possible for this to be true of RlySm (under 9 students) or RlyLg (over 38 students) simply because each of these only came up in one out of the six analyses in this study. Many (though not all) of the corresponding slopes also had small magnitudes, so even if there was a pattern it would not indicate much of a practical effect on Grades. It is worth noting, though, that the slopes corresponding to Lit and Med were mostly (5 out of 6 analyses) negative, with two of those on Med being statistically significant. At the same time, the slopes corresponding to Sm were mostly (4 out of 5 analyses that involved DLs which fell into this category) positive, with one of these being statistically significant. The slopes corresponding to the continuous Lecture size predictor variable are very evenly split between positive and negative, though the two that are statistically significant are both negative. Also, while the magnitudes of these slopes may seem small, most of them, including the two statistically significant ones, are actually substantial given the range of Lecture sizes in the data. For instance, in 7A during the summer, where this slope is statistically significant, the minimum Lecture size is 84 students while the maximum Lecture size is 145 students. This means that, after accounting for the effects of all other predictor variables, a student in 7A during the summer whose Lecture had the maximum number of students received a Grade that was on average 1.20 points (on a 4.00 scale) lower than that of a student whose 7A summer Lecture had the minimum number of students.

Altogether, this seems to indicate that Lecture, and especially DL, sizes do not have much impact on student understanding of physics concepts in the types of courses were part of this study (at least to a point, since a DL of say, 100 students, would be a different story that is beyond the scope of this study given the DL sizes that were part of the sample data). Therefore, in light of this study, there is less reason to focus attention or resources on reducing the size of introductory university physics classes than there otherwise might have been. However, there are some caveats to this.

First off, while the slopes associated with Lecture size yielded mixed results, it is intriguing that two of them are statistically significant with rather large magnitudes (in the context of Lecture size ranges), and that both of these are negative. It is also relevant to note that both of these are related to summer Lectures where Lecture size means the direct size of lecture sections, unlike during the regular academic year where each Lecture includes two lecture sections (taught back-to-back by the same Lecture instructor), and so Lecture size is not directly tied to the size of any given lecture section (though it is still strongly related). Note, though, that 7A and 7B during the summer are also the situations where the later Lecture start time, or more accurately, the DL start times that correspond exactly to the later Lecture start time, had large and statistically significant slopes. Therefore, it is possible that the apparent effect that Lecture size has on Grade here is related to the effect that these start times have on Grade through some additional unknown factor(s) that are associated with DL start times, Lecture start times, and Lecture sizes in 7A and 7B during the summer.

When it comes to DL sizes, in addition to the issues discussed in "Limitations and Future Work," the biggest confounding factor here is really the fact that a very small portion of the variance in Grades exists at the DL level to begin with (and the uncertainty in this variance, as measured by standard errors in it, is quite high compared to the variance itself). This means that level 2 predictor variables, including DL size predictor variables, should not have much of an impact on Grades regardless of anything else since the primary purpose of predictor variables is to explain variance in the outcome variable and predictor variables at a given level should, at least theoretically, only explain variance at that level.

In terms of why there is so little variance in Grade between DLs (within a given Lecture), perhaps this means that DLs are so well organized, planned, and standardized that the differences between them do not have much of an impact on student learning. Another possibility, though, is that quizzes and exams, which are typically created and administered by Lecture instructors during lecture sections and are the primary determinant of Grades, may not necessarily fully or adequately reflect what is being taught in DLs (essentially, maybe the assumption that they do, which is discussed further in Appendix B as part of a discussion around content validity, does not hold as well as it was initially believed to). There is also a possibility that students do most of their learning for the courses that were part of this study during lecture sections (even though that is not the intention of these courses), or even outside of class-time and class-based learning (which seems more plausible since, regardless of a course's class-based format, students often spend a lot of time studying).

Finally, it is possible that student learning in the courses that were part of this study is not that strongly influenced by class environments or instructor guidance (in other words, maybe the assumption of malleability, which is discussed further in Appendix B as part of the discussion around content validity, does not hold as well as it was initially believed to). In particular when it comes to class size, one of the principles behind these courses is for students to engage in peer learning, especially in their groups during DL while the TA plays a more supportive, guiding role. Since DL size likely does not have a major impact on the size, composition, or functioning of these groups, even if it does often have a major impact on the number of groups in a given DL, perhaps student learning in DLs is not that strongly influenced by DL size because most learning during DL occurs between students in a given group regardless of their DL's size. Even in such a case, though, there is a question of how the DL and Lecture size expectations in the courses that were part of this study effect the structure and format of these courses. More specifically, it is unknown whether or not there might be a structure or format that better facilitates student learning, but where implementing it would require different expectations around the typical sizes of DLs and/or Lectures. For instance, smaller class sizes could make it possible for quizzes and exams to include more conceptual questions, which tend to take longer to grade than calculational problems do, but are also usually better at gauging physics understanding than standard calculational problems or the memorization and repetition that frequently accompanies them [9]. Smaller class sizes could also make it more reasonable to let students take multiple assessments of the same topic in such a way that more recent assessments replace previous ones in order to demonstrate proficiency even if their initial understanding was lacking [14]. The above questions would all benefit from further study.

Limitations and Future Work

Appendix B discusses some limitations of this study (and some potential future work) around the nature and meaning of grades (like questions about what material from the courses that were part of this study Grades are based on), as well as possible alternative outcome variables (like final exam scores or scores on the FCI - the Force Concept Inventory). The skewness and ceiling and floor effects that are present in some of this study's continuous variables, as described in "Levels and Outcome Variable" and "Analysis Format and Summary Data," also present some clear limitations to this study. Besides these issues and the questions raised in "Slope Discussion and Overall Conclusions," one obvious limitation of this study is the missing observations. As discussed in "Analysis Format and Summary Data," due to these missing data points coming from missing GPAs due to fewer than 12 Units being associated with such observations, this missing data systematically excludes first quarter freshmen, transfer students, and graduate students, along with students who do not primarily attend the University. However, beyond these groups, it is assumed that the rest of the missing observations are random and that they do not systematically exclude or disproportionately impact the results from any other groups of students who take the courses that were part of this study. Another assumption of this study is that the exact definition of DL size categories does not make a significant difference for the results (and thus, the conclusions). Future studies could conduct analyses using a few slightly different DL size category definitions (where the categories are shifted a bit and/or the number of students in each category is maybe 5 or 7 instead of 6) and check for general consistency between corresponding analyses where the only difference is in these definitions.

There are also several additional factors that may have an impact on Grade and could potentially be controlled for in future studies (provided access to the proper data sets is granted). These factors could be addressed through either new predictor variables and/or a new random effects/levels structure. For instance, including a continuous predictor variable for students' high school GPAs, either along with, or even instead of, overall University GPA, could help account for their general prior knowledge and understanding. Similarly, including continuous predictor variables for students' prior grades in the University's introductory calculus or chemistry courses, either along with, or even instead of, overall University GPA, could help account for their prior knowledge and understanding of topics that are related to those covered by the courses in this study. Including students' socioeconomic status as a predictor variable (either a continuous one or a set of categorical ones) would be a good way to account for some of the impacts that this aspect of student backgrounds and structural barriers have on grades, the same way that including student gender (which should have more than two identified options), race and ethnicity, and U.S. citizenship status account for some of the impacts that other aspects of student backgrounds and structural barriers have on grades.

Another major factor to potentially control for in future studies is a DL's TA, especially since most of the class-time and class-based learning in the courses that were part of this study is done during DL and TAs often have their own unique styles (which has a direct effect, as well as an indirect effect since students sometimes try to get into a DL with a certain TA). Given the large number of TAs in the data, it probably would not be reasonable to account for their effect through a set of categorical predictor variables (though if it was, these predictor variables would be at the DL level). Instead, the best way to account for the effect of TAs would be to treat them as level 3 such that DLs are nested within TAs (since most TAs taught multiple DLs during the years that this study covers). However, this would not be possible while simultaneously keeping the Lecture level, not only because having four levels would be difficult, but also because each Lecture has multiple TAs and, during the years that this study covers, most TAs taught DLs that were associated with multiple Lectures. This means that Lectures would not be nested within TAs nor would TAs be nested within Lectures. Despite this, since Lectures have such a substantial effect on Grade, it would still be desirable to account for their effect in some way and there are a few options for accomplishing this. One possibility would be to not include a Lecture level but to mean-center Grades at the Lecture level and use this as the outcome variable [10, p. 59-69]. In other words, the Grades in each Lecture could be converted to z-scores relative to that Lecture (z-scores which are subsequently used as the outcome variable) by subtracting the mean Grade in each Lecture from each individual Grade in that Lecture and then dividing the results by that Lecture’s Grade standard deviation. Another possibility would be using more advanced HLM techniques that do not require a strict nesting structure, though doing so in this case would also involve using a four level model [10, p. 171-187].

One last possibility to account for the effect of TAs would be to conduct a quasi-experiment where multiple TAs (who, during the regular academic year, usually teach two different DL sections per quarter that are both within the same Lecture) are assigned one DL section with the standard number of students (27-32) and one DL section with substantially more or fewer students (such that it falls into the Small, Literature, Medium, or Large category, since the Really Small and Really Large categories naturally arise so rarely that they do not warrant as much study). Comparisons could then be made between corresponding pairs of DL sections, while controlling for as many other factors as possible. Such an experiment would not only control for the effect of TAs, but could also help improve statistical power (i.e. make it easier to find relationships between DL size and Grades if these relationships exist in the population) by addressing the relatively low number of observations that naturally occur in all of the DL size categories besides the standard one. A quasi-experiment would also help mitigate the broader issue of students self-selecting into certain DLs or Lecturers, which directly affects the sizes of different DLs and Lecturers, as well as the types of students in different DLs and Lectures (for instance, students with more Units, which allows them to register for classes earlier than students with fewer Units, may disproportionately end up in DLs with more desirable start times).

Another idea that would be interesting to explore in the future is treating Lecture instructor as a level in order to compare the results of such analyses to the results obtained by treating Lecture instructors as a set of Lecture level categorical predictor variables, as was done in this study. Here it is relevant to note that an earlier version of the analysis presented in this dissertation was conducted using only two levels; students (or observations) and DLs, while omitting the Lecture level. This previous analysis found a substantial amount of variation in Grades (i.e. an ICC greater than 0.05) at the DL level, but it is now clear that most of that variation was actually attributable to the Lecture level. It is therefore possible that if a Lecture instructor level is included, it may turn out that most of the variation in Grades at the Lecture level in this study really exists at the Lecture instructor level. Treating Lecture instructor as its own level would also more easily facilitate the inclusion of predictor variables that are related to the Lecture instructor, as opposed to the Lecture itself, like their gender or teaching style, if there is interest in studying the effects of such things on student Grades or other outcome variables (as opposed to combining these effects into the singular effect of individual Lecture instructors). However, this type of analysis would involve either making Lecture instructor a fourth level that Lecturers are nested within and/or removing the DL or Lecture level. One more idea along these lines could be using more advanced HLM techniques to treat the true lecture as a level rather than treating the combined Lecture as a level (where, during the regular academic year, a Lecture includes both of the lecture sections that are taught back-to-back by the same instructor).

Similarly, it would also be interesting to explore what happens when observations are treated as a separate level from students in order to compare the results of such analyses to the results obtained by treating the observation level as effectively equivalent to the student level, as was done in this study. When treated as distinct from observations, students are not strictly nested within DLs or Lectures due to some students taking the same course multiple times (Repeats) and also due to students often taking multiple courses in the sequence of three courses that were part of this study (where, unlike in the study presented here, this type of analysis could potentially combine data from all three of these courses). Therefore, this type of analysis would require more advanced HLM techniques (and likely more than three levels).

One potential future study that would be related to, but distinct from (as opposed to a modification of) this study, would involve logistic or logit analyses around what factors are associated with whether or not a student dropped one of the Lectures in this study [10, p. 112-140]. Finally, regardless of what outcome variable(s), predictor variables, and random effects structure(s) are used in potential future studies, one last possibility for quantitatively studying the effect of class size on cognitive outcomes (those related to student achievement and understanding of the material) would be to incorporate random slopes and/or interactions (since these things were not included in this study). Doing so would also need to involve robust theoretical work to determine which predictor variable(s) should have random slopes and what interaction(s) to include.

Beyond quantitative studies, it would also help answer this study's research question if more qualitative work is done on the role of class size in the types of courses that were part of this study (using techniques like interviews with students, TAs, and Lecture instructors; DL and Lecture observations; open ended survey questions to students, TAs, and Lecture instructors; etc.). Similarly, beyond studies of class size and cognitive outcomes, it would be helpful if some studies were conducted on non-cognitive outcomes (those which are not directly tied to student achievement) in the types of courses that were part of this study, whether qualitatively or quantitatively (such as by using survey prompts that incorporate Likert scales, like strongly agree, agree, disagree, etc.) [15]. For instance, a few relevant non-cognitive research questions might be: What is the relationship between DL size and how DLs operate; how do students, TAs, and Lecture instructors feel about different class sizes and the effect they have on student learning; and how do different class sizes effect the workload, as well as the quality of grading and other work, for TAs and Lecture instructors?

Lastly, in order to better gauge what population the findings of this and related studies apply to, it would be fruitful to conduct a version of this study and/or any of those described above on other courses. These could include other introductory physics courses at the University, different types of introductory chemistry or calculus courses at the University, or different types of introductory physics, chemistry, or calculus courses at other colleges or universities (or even high schools). Note that here, "different types" of courses means courses that are conducted and formatted in different types of ways, with a focus on potential differences between active-learning based courses and more traditional ones.

References

[1] Achilles, C. M. (2012). *Class-Size Policy: The STAR Experiment and Related Class-Size Studies* (2nd ed., Vol. 1, Issue brief) (C. D. Shiffman, Ed.). Tecumseh, MI: National Council of Professors of Educational Administration (NCPEA). (ERIC Document Reproduction Service No. ED540485)

[2] Goldstein, H., Yang, M., Omar, R., Turner, R., & Thompson, S. (2000). Meta-analysis using multilevel models with an application to the study of class size effects.

[3] Wyss, V. L., Tai, R. H., & Sadler, P. M. (2007). High school Class-size and College Performance in Science.

[4] Leone, C. D., Potter, W. H., Ishikawa, C. M., Blickenstaff, J., & Hession, P. L. (2001). Class Size Effects in Active Learning Physics Courses.

[5] Bettinger, E. P., & Long, B. T. (2017). Mass Instruction or Higher Learning? The Impact of College Class Size on Student Retention and Graduation. *Education Finance and Policy*, *13*(1), winter 2018, 97-118.

[6] Mckagan, S. B., Perkins, K. K., & Wieman, C. E. (2007). Reforming a large lecture modern physics course for engineering majors using a PER-based design.

[7] Wieman, C. (2007). A new model for post-secondary education, the Optimized University.

[8] Gee, K. A., & Wong, K. K. (2012). A cross national examination of inquiry and its relationship to student performance in science: Evidence from the Program for International Student Assessment (PISA) 2006.

[10] Hox, J. J. (2010). *Multilevel Analysis: Techniques and Applications*. New York, NY: Routledge.

[9] Ashbaugh West, E. L. (2009). *Identifying the elements of physics courses that impact student learning: Curriculum, instructor, peers, and assessment*. Ann Arbor, MI: ProQuest LLC. Retrieved January 05, 2022, from http://www.proquest.com/en-US/products/dissertations/individuals.shtml

[11] Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and Data Analysis Methods* (Second, Vol. 1, Ser. Advanced Quantitative Techniques in the Social Sciences). Sage Publications.

[12] Theobald, E. (2018). Students Are Rarely Independent: When, Why, and How to Use Random Effects in Discipline-Based Education Research. *CBE—Life Sciences Education*, *17*(3), 1–12. doi:10.1187/cbe.17-12-0280

[13] Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., . . . Welsh, M. E. (2016). A Century of Grading Research: Meaning and Value in the Most Common Educational Measure. *Review of Educational Research, 86*(4), 803-848. doi:10.3102/0034654316672069

[14] Webb, D. J., Paul, C. A., & Chessey, M. K. (2020). Relative impacts of different grade scales on student success in introductory physics. *Physical Review Physics Education Research, 16*(2), https://doi.org/10.1103/PhysRevPhysEducRes.16.020114

[15] Dee, T., & West, M. (2011). The Non-Cognitive Returns to Class Size. *Educational Evaluation and Policy Analysis*, *33*(1), 23–46. doi:10.3102/0162373710392370

[16] Bandalos, D. L. (2018). *Measurement Theory and Applications for the Social Sciences*. New York: Guilford Press.

[17] DeBoer, G. E. (2000). Scientific Literacy: Another Look at Its Historical and Contemporary Meanings and Its Relationship to Science Education Reform. *Journal of Research in Science Teaching, 37*(6), 582-601. doi:10.1002/1098-2736(200008)37:63.0.CO;2-L

[18] Doughty, L., & Caballero, M. D. (2014, July). Rubric Design for Separating the Roles of Open-Ended Assessments. In *PERC 2014 Proceedings*. Retrieved from https://arxiv.org/pdf/1407.3297.pdf

[19] Paul, C., Potter, W., & Weiss, B. (2014). Grading by Response Category: A simple method for providing students with meaningful feedback on exams in large courses. *The Physics Teacher, 52*(8), 485-488. doi:10.1119/1.4897587

[20] Lin, S., & Singh, C. (2013). Can free-response questions be approximated by multiple-choice equivalents? *American Journal of Physics, 81*(8), 624-629. doi:10.1119/1.4804194

Appendices

Appendix A: General Three Level Basic HLM Equations

Assuming the outcome variable is continuous and at level 1, the general equations for a basic three level HLM model can be written as:

Level 1:

$$\mathrm{OutcomeVariable}_{ijk} = \beta_{0jk} + \beta_{1jk} * \mathrm{Level1Variable1}_{ijk} + \ . \ . \ . \ + \beta_{Mjk} * \mathrm{Level1VariableM}_{ijk} + \varepsilon_{ijk}$$

Level 2:

$$\beta_{mjk} = \gamma_{m0k} + \gamma_{m1k} * \mathrm{Level2Variable1}_{jk} + \ . \ . \ . \ + \gamma_{mNk} * \mathrm{Level2VariableN}_{jk} + u_{mjk}$$

$$\forall m \in \{0, \ . \ . \ ., \ M\}$$

Level 3:

$$\gamma_{mnk} = \pi_{mn0} + \pi_{mn1} * \mathrm{Level3Variable1}_{k} + \ . \ . \ . \ + \pi_{mnP} * \mathrm{Level3VariableP}_{k} + \nu_{mnk}$$

$$\forall m \in \{0, \ . \ . \ ., \ M\}, \forall n \in \{0, \ . \ . \ ., \ N\}$$

Where i is an index labeling the entities that level 1 data is coming from (such as students in the Students/Classes/Schools example), j is an index labeling the entities that level 2 data is coming from (such as classes in the Students/Classes/Schools example), and k is an index labeling the entities that level 3 data is coming from (such as schools or universities in the Students/Classes/Schools example). Similarly, m is an index labeling level 1 predictor variables, n is an index labeling level 2 predictor variables, and p is an index labeling level 3 predictor variables, which implicitly appears in the above equations through terms of the form $\pi_{mnp} * \mathrm{Level3Variablep}_{k}$. There are M level 1 predictor variables, N level 2 predictor variables, and P level 3 predictor variables. Note that the three level model has been specified because reference materials (including [10], which much of this Appendix is based on) often focus on the two level model but, in the author's experience, it is not always clear how to extend the two level model to three or more levels, whereas extending the three level model to more than three levels is, at least theoretically, fairly straightforward (plus, as discussed at the end of the "HLM" portion of "Statistical Techniques and Random Effects," it is rare to apply HLM to more than three levels).

Also note that these equations are written on a level-by-level basis. It is possible to write a composite description of HLM equations by taking the equations for intercepts and slopes (i.e. all equations at levels beyond level 1) and plugging them into the corresponding terms in lower level equations, but doing so can get messy while not being all that informative when the model that is under consideration includes few, if any, interaction terms, so this paper will be sticking to the level-by-level description.

It is important to realize that these models can get quite complicated rather quickly (especially with more than two levels), so while the descriptions presented in this Appendix are completely general, in practice, it is rare for these models to include all possible terms in their respective equations. Instead, theory and empirical evidence (frequently in the form of testing a series of increasingly more complex models while dropping non-significant terms along the way), along with philosophical notions like a desire to have the simplest possible model that still makes sufficiently good predictions, are used to decide which terms to include and which ones to leave out.

On a related note, standard regression techniques tend to fit data based on ordinary least squares (OLS) procedures where the best fit line is determined by minimizing the sum of the squares of the residuals (error terms). HLM, however, typically uses maximum likelihood estimation (MLE) where the best fit line is determined by maximizing some likelihood (probability) function (the details of which are beyond the scope of this paper). MLE can also be used to do standard regression whereas OLS cannot be used in HLM without substantial modifications because in HLM there are different types of error terms at different levels and how to appropriately weight and properly use each of them is not well-defined unless explicitly specified.

Intercepts and Regression Coefficients

In the above equations, intercepts $\left(\beta_{0jk}, \gamma_{00k}, \text{and } \pi_{000}\right)$ represent the average value of the outcome variable when all predictor variables take on a value of 0 (either for a given jth level 2 cluster within a given kth level 3 cluster, a given kth level 3 cluster, or overall if there is no j or k subscript), making them intimately tied to error terms (the difference between actual and predicted values of the outcome variable). When it comes to $\beta_{0jk}, \gamma_{00k}, \text{and } \pi_{000}$, the "prediction" here would be for a hypothetical observation where all predictor variables take on a value of 0.

Slope coefficients (either for a given jth level 2 cluster within a given kth level 3 cluster, a given kth level 3 cluster, or overall if there is no j or k subscript) on dummy predictor variables (predictor variables that are binary and have a value of either 0 or 1) represent the average amount by which the outcome variable is different for those who have the characteristic that is assigned a value of "1" compared to those who have the characteristic that is assigned a value of "0." Slope coefficients (either for a given jth level 2 cluster within a given kth level 3 cluster, a given kth level 3 cluster, or overall if there is no j or k subscript) on predictor variables which are part of a set of categorical predictor variables (a set of complete and mutually exclusive dummy variables where every observation is assigned a value of 1 for exactly one such variable, typically the one associated with the name of the variable, and 0 for all of the others) represent the average amount by which the outcome variable is different for those who have the corresponding characteristic compared to those who are part of the reference category (a chosen member of the compete and mutually exclusive set) after controlling for all other predictor variables in the model.

Slope coefficients (again, either for a given jth level 2 cluster within a given kth level 3 cluster, a given kth level 3 cluster, or overall if there is no subscript) on continuous predictor variables represent the average amount by which the outcome variable changes when the value of said variable increases by 1 unit after controlling for the effect of all other predictor variables in the model.

For example, in the Students/Classes/Schools example, when all predictor variables have a value of 0, β_{035} is the average value of exam score within class 3 which is itself within school or university 5, γ_{005} is the average exam score within school or university 5 (accounting for all classes in school or university 5), and π_{000} is the average exam score in the overall study (accounting for all classes and schools or universities).

Now say that there is a level 1 dummy variable in the Students/Classes/Schools example called Clubs which is labeled as level 1 predictor variable 4 and has a value of 1 for those who participate in extracurricular activities and a value of 0 for those who do not. In this case, β_{435} is the average difference in exam score for those who participate in such activities compared to those who do not in class 3 within school or university 5 after accounting for the effect of all other predictor variables.

If, in the Students/Classes/Schools example, the level 1 continuous predictor variable PriorGrade (a given student's average grade on prior exams) is labeled as level 1 predictor variable 2, then β_{235} is the average change in exam score due to a 1 unit (likely 1 point) change in PriorGrade for students in class 3 within school or university 5 after accounting for the effect of all other predictor variables. Now suppose there is a level 2 continuous predictor variable TeachYrExp (the teacher's temporal experience in years). If TeachYrExp is labeled as level 2 predictor variable 1, then γ_{015} is the average change in exam score due to a 1 unit (1 year) change in TeachYrExp for students in school or university 5 after accounting for the effect of all other predictor variables. Finally, say there is a level 3 continuous predictor variable Mean_Income (the mean parental income of students in a given school or university). If Mean_Income is labeled as level 3 predictor variable 4, then π_{004} is the average change in exam score due to a 1 unit (likely 1 dollar) change in Mean_Income for all after accounting for the effect of all other predictor variables.

Note that in an HLM context, interactions between predictor variables (which arise due to predictor variables either moderating or mediating each other) are addressed slightly differently depending on which level(s) these predictor variables are associated with. When they are at the same level, interactions are accounted for by defining a new predictor variable at that level (with its own corresponding slope at that level) such that this new variable is the product of the two interacting predictor variables. However, interactions between predictor variables at different levels are addressed by treating lower level slopes as functions of higher level predictor variables, meaning the slope on such an interaction will ultimately come from the highest level that is part of that interaction but will do so through this functional relationship after substituting higher level equations into lower level slopes (though the interpretation of what an interaction means in terms of its effect on the outcome variable is the same as it is for interactions in standard regression analyses).

For instance, in the Students/Classes/Schools example, the interaction between PriorGrade and TeachYrExp within school or university 5 would have a slope γ_{215} which comes from taking the level 1 slope on PriorGrade $\left(\beta_{2j5} \text{ for class j}\right)$ and substituting in the level 2 equation for β_{2j5} which includes (among other things) both γ_{205} (which is related to β_{2j5} through error terms) and $\gamma_{215} * \text{TeachYrExp}$ meaning there is now a term $\gamma_{215} * \text{TeachYrExp} * \text{PriorGrade}$. Similarly, the interaction between TeachYrExp and Mean_Income would have a slope π_{014} which comes from taking the level 1 intercept (β_{0jk} for class j and school or university k), substituting in the level 2 equation for β_{0jk} which includes (among other things) $\gamma_{01k} * \text{TeachYrExp}$ for school or university k, and then substituting in the level 3 equation for γ_{01k} which includes (among other things) both π_{010} (which is related to γ_{01k} and β_{0jk} through error terms) and $\pi_{014} * \text{TeachYrExp}$ meaning there is now a term $\pi_{014} * \text{Mean\_Income} * \text{TeachYrEx}$. Finally, the interaction between PriorGrade and Mean_Income would have a slope π_{204} which comes from taking the level 1 slope on PriorGrade (β_{2jk} for class j and school or university k), substituting in the level 2 equation for β_{2jk} which includes (among other things) γ_{20k} for school or university k, and then substituting in the level 3 equation for γ_{20k} which includes (among other things) both π_{200} (which is related to γ_{20k} and β_{2jk} through error terms) and $\pi_{204} * \text{Mean\_Income}$ meaning there is now a term $\pi_{204} * \text{Mean\_Income} * \text{PriorGrade}$.

Interactions between predictor variables at more than two different levels can be systematically worked through in this type of manner as well, though including interactions between more than two predictor variables is rare in any regression technique (HLM or otherwise).

Error Terms

One aspect of the above equations that may seem particularly unfamiliar are the terms ε_{ijk}, u_{mjk}, and ν_{mnk}. These are error terms (with the latter two forming the mathematical basis for random effects). When m and n are both 0, these error terms are the difference between actual values of the outcome variable and the corresponding predicted (from the regression equations) values at each of the three levels. When m and/or n are not 0, u_{mjk} and ν_{mnk} are the difference between regression coefficients (slopes) for different level 2 or level 3 clusters after accounting for the (fixed) effect of all predictor variables (where said slopes are associated with either a given level 1 predictor variable, a given level 2 predictor variable, or an interaction term between a given level 1 predictor variable and a given level 2 predictor variable).

Getting a bit more specific, the residual error term ε_{ijk} might seem somewhat familiar since it is similar to the error term in standard regression analyses, except instead of being the difference between the actual value of the outcome variable for a given observation and the overall predicted value of the outcome variable for that observation, it is the difference between the actual value of the outcome variable for a given observation (labeled by i) and the predicted value of the outcome variable for that observation within a given (j^{th}) level 2 cluster (which is itself nested within a given, or k^{th}, level 3 cluster). Note that residual error terms within a given level 2 cluster are assumed to follow a Normal Distribution and homoscedasticity.

u_{0jk} is the difference between the actual average value of the outcome variable for a given j^{th} level 2 cluster (within a given k^{th} level 3 cluster) and the overall predicted value of the outcome variable in the j^{th} level 2 cluster. Similarly, ν_{00k} is the difference between the actual average value of the outcome variable for a given k^{th} level 3 cluster and the overall predicted value of the outcome variable in the k^{th} level 3 cluster.

For example, in the Students/Classes/Schools example, analyzing the outcome variable (exam score) across multiple classes and schools or universities will yield a prediction for each student's grade. In standard regression techniques, the error term is simply the difference between a given student's actual exam score and their predicted exam score, but in HLM there is a prediction for a given student who is in a given class which is in a given school or university. In this situation, ν_{00k} is the difference between the actual average exam score in a given school or university and the overall predicted exam score for that school or university, u_{0jk} is the difference between the average exam score in a given class in a given school or university and the overall predicted exam score for that class in that school or university, and ε_{ijk} is the difference between a given student's actual exam score and the predicted exam score for that student in that class in that school or university.

When m is not 0, u_{mjk} represents the difference between the slope for the mth level 1 predictor variable in the jth level 2 cluster within the kth level 3 cluster (i.e. β_{mjk}) and the overall slope for the mth level 1 predictor variable in the kth level 3 cluster after accounting for all level 2 predictor variables. In the Students/Classes/Schools example, say that in school or university 4 (so the 4th level 3 cluster) there is an overall slope (across all classes) on the level 1 predictor variable PriorGrade (a given student's average grade on prior exams). There is also a slope on PriorGrade for class 6 within school or university 4 which may be different than the overall slope for school or university 4. The difference between these two slopes would be u_{264} (assuming PriorGrade is labeled as the second level 1 predictor variable) when the values of all class level variables are held constant.

When m is 0 but n is not 0, ν_{0nk} represents the difference between the slope for the nth level 2 predictor variable in the kth level 3 cluster and the overall slope for the nth level 2 predictor variable after accounting for all level 3 predictor variables. In the Students/Classes/Schools example, say there is an overall (across all schools or universities) slope on the level 2 predictor variable TeachYrExp (the teacher's temporal experience in years). There is also a slope on TeachYrExp for school 4 which may be different than the overall slope. The difference between these two slopes would be ν_{014} (assuming TeachYrExp is labeled as the first level 2 predictor variable) when the values of all school level variables are held constant.

When neither m nor n is 0, ν_{mnk} represents the difference between the slope for the interaction term between the m$^{\text{th}}$ level 1 predictor variable and the n$^{\text{th}}$ level 2 predictor variable in the k$^{\text{th}}$ level 3 cluster and the overall regression coefficient for the n$^{\text{th}}$ level 2 predictor variable after accounting for all level 3 predictor variables. In the Students/Classes/Schools example, say we posit that there is an interaction between the level 1 predictor variable PriorGrade and the level 2 predictor variable TeachYrExp (maybe teachers with more years of experience grade things differently than those with fewer years of experience). Then there is an overall (across all classes and schools or universities) slope on this interaction term in the model. There is also a slope on this interaction term for school or university 4 which may be different than the overall slope. The difference between these two slopes would be ν_{314} when the values of all school or university level variables are held constant.

Error Term Variance and Intraclass Correlation Coefficients

As with standard regression techniques, it is possible to find the variances of, and covariances between, these error terms. The variance in $\varepsilon_{\mathrm{ijk}}$ (known as the residual variance and denoted $\sigma_{\varepsilon}{}^{2}$) represents the variation in the outcome variable within a given level 2 cluster (which again, is itself nested within a given level 3 cluster) after accounting for the effects of all predictor variables. Unless stated and specified otherwise, it is assumed that this variance follows a Normal distribution with a mean of 0 and that it is the same for all level 2 clusters. The variance in u_{0jk} ($\sigma_{u0}{}^2$) represents the variation in the outcome variable between level 2 clusters within a given level 3 cluster after accounting for the effect of all level 2 predictor variables in the equation for the level 1 intercept (β_{0jk}). The variance in ν_{00k} ($\sigma_{\nu00}{}^2$) represents the variation in the outcome variable between level 3 clusters after accounting for the effect of all level 3 predictor variables in the equation for the level 2 intercept (γ_{00k}).

The variance in u_{mjk} when m is not 0 ($\sigma_{um}{}^2$) represents the variation in the error term u_{mjk} described above. Similarly, the variance in ν_{0nk} when n is not 0 and ν_{mnk} when neither m nor n is 0 ($\sigma_{\nu0n}{}^2$ and $\sigma_{\nu mn}{}^2$ respectively) represent the variation in their respective error terms which were also described previously. $\sigma_\varepsilon{}^2$, $\sigma_{u0}{}^2$, and $\sigma_{\nu00}{}^2$ can be defined in this manner as well, but it is informative to be more explicit about these, which is why they were discussed separately. One reason for this is because they can be used to determine intraclass correlation coefficients (ICCs) for levels 2 and 3, which are defined (respectively) as:

$$\mathrm{ICC}_2 = \rho_2 = \frac{\sigma_{u0}{}^2}{\sigma_\varepsilon{}^2 + \sigma_{u0}{}^2 + \sigma_{\nu00}{}^2}$$

$$\mathrm{ICC}_3 = \rho_3 = \frac{\sigma_{\nu00}{}^2}{\sigma_\varepsilon{}^2 + \sigma_{u0}{}^2 + \sigma_{\nu00}{}^2}$$

The ICC for a given level is the fraction or proportion (which can be converted to a percentage if desired) of total variance (total variance before any predictor variables are accounted for) in the outcome variable that is attributable to variation between clusters (as opposed to variations between observations within a given cluster) at that level [10, p. 34]. Therefore, ICC_2 is the proportion of total variance in the outcome variable that is due to variations between level 2 clusters in a given level 3 cluster while ICC_3 is the proportion of total variance in the outcome variable that is due to variations between level 3 clusters. For instance, in the Students/Classes/Schools example, the total variance in exam scores can be partially attributed to the variation between students in a given class in a given school or university ($\sigma_{\varepsilon}{}^{2}$), partially attributed to the variation between classes in a given school ($\sigma_{u0}{}^{2}$), and partially attributed to the variation between schools ($\sigma_{\nu00}{}^{2}$).

It is possible to define ICC_1 as the proportion of variance in the outcome variable that is due to variations between observations in a given level 2 cluster (which is itself nested within a given level 3 cluster), but this is generally not done, primarily because it is often more important to know what proportion of the variance exists between level 2 clusters within a given level 3 cluster, as well as between level 3 clusters, than it is to know what proportion of the variance exists within a given level 2 cluster. Secondarily, defining ICC_1 would be redundant and unnecessary since $\mathrm{ICC}_1 + \mathrm{ICC}_2 + \mathrm{ICC}_3 = 1$. The fact that variance can be parsed out and attributed to different types of clusters in this way is yet another example of useful information being garnered from Multilevel Modeling that cannot be determined through standard regression techniques. In HLM, the first model that is typically analyzed is the Null Model which does not include any predictor variables since its objective is to figure out how much variance exists at different levels before any of this variance is explained by (and therefore, reduced as a result of) predictor variables (at one or more levels where, theoretically and conceptually, predictor variables that are associated with a given level should primarily or exclusively explain and reduce variance that is associated with that level, but not explain or reduce variance that is associated with other levels).

One final role that the ICC plays is helping to determine if a given type of clustering should be treated using random effects, as opposed to through a set of categorical predictor variables (or even just not incorporating such clustering into the analysis at all). An informal rule says that if the Null Model ICC associated with a given type of clustering (such as classes or schools or universities in the Students/Classes/Schools example) is at least 0.05 (5%), then that type of clustering is a good candidate for a Multilevel Modeling approach (i.e. random effects).

Appendix B: Grades in the Context of Measurement Theory (in General and in Relation to This Study)

Validity and Bias

Considering the outcome variable in this study is the overall Lecture Grades for individual students, it is clear that the nature, meaning, legitimacy, and interpretation of grades is foundational to this study. Questions around these issues are themselves questions about the validity of grades based on the most widely accepted definition of the term, which is essentially the degree to which a measurement device ("test") actually measures whatever underlying construct (such as knowledge of a certain topic or level of satisfaction with a service) it is meant to measure, though it should be noted that there is still no fully agreed upon definition of measurement validity [16, p. 255]. One common form of validity is known as criterion-related validity and is based on a test's relationship to (which often means its results' correlation with the results of) another known test that measures the same latent trait (underlying construct). The known test can either be done at the same time (known as concurrent validity) or in the future (known as predictive validity) [16, p. 257].

One frequently held belief is that grades are "what students 'earn' for their achievement" [13]. This assertion assumes that grades are based solely on students' understanding of the relevant subject matter which, if true, would be useful to this study since the goal here is to explain and predict students' understanding of physics principles (the latent trait of interest in this study). Such understanding is an important construct for three primary reasons. First, people tend to be curious about the fundamental laws of nature and studying physics is a good way of learning about such things. Secondly, the problem solving skills required to analyze physics problems are applicable to other scenarios as well and finally, physics concepts are directly useful in a variety of ways, from everyday situations, like driving, to more career specific pursuits, like engineering. Furthermore, in addition to students who take physics courses being concerned about their physics grades because of the way that grades are used to sort people and determine who has access to certain opportunities (like internships and medical school), many of them probably have at least some curiosity about the laws of physics, both in general and in terms of how physics relates to their chosen fields and majors (so for the most part when it comes to the courses involved in this study, biological applications of physics concepts). Since it is not possible to directly observe someone's understanding of physics, this study took overall Lecture Grades to be a test (in the generalized sense) that measures such things [16, p. 3].

However, despite some commonly held beliefs, there are a range of problems with the assumption that grades are exclusively, or even largely, a measure of academic achievement. For one, in K-12 schooling, numerous studies of criterion-related validity have been conducted to examine the relationship between grades (sometimes aggregate grades, like GPAs, and sometimes grades in specific classes, like math) and outcomes on so-called "achievement" or "intelligence" tests. These studies have consistently demonstrated a moderate relationship between the two (even as the composition of such tests, and the educational system more broadly, have gone through substantial changes over the years), implying a significant but modest relationship between grades and achievement (or "intelligence") as defined by such tests (the criterion in this case) [13]. It should be noted that these sorts of studies assume that such tests actually measure achievement (or "intelligence") and going even deeper, that the constructs of "achievement" and "intelligence" have well-defined meanings (i.e. that these tests are themselves valid). While university courses are obviously different than K-12 schooling, it is likely that similar relationships exist there as well, though it would be beneficial to directly make this determination through empirical studies. For instance, one possible format for studies of this type could be to determine the relationship between student grades in introductory college physics courses that teach Newton's Laws and student scores on the Force Concept Inventory (FCI), a well-known and commonly used assessment of conceptual (though not calculational) understanding of forces and Newton's Laws. This could be especially helpful considering there is reason to believe that conceptual, explanation-based questions are a better gauge of physics understanding (and thus, a more valid measure of this underlying construct) than questions centered around calculations, or even diagrams [9]. Since the FCI is multiple choice, scores on it would also not be subject to concerns around the consistency of different graders, as discussed in the context of reliability below. In fact, some work along these lines that could be built upon has already been done, including some analysis involving introductory physics courses at the University (but ones that are taken primarily by engineering majors, as opposed to the courses that were part of this study, which are taken primarily by bioscience majors) [9]. In the courses that were part of this study, this type of analysis would also have the added benefit of helping to determine the role that Lecture instructors (and Lectures more broadly) play in student learning by separating assessments of student learning from Grades, which Lecture instructors in these courses influence through both their teaching and their logistical and administrative roles, as discussed in "Site, Sample, and Population," among other places.

Building off of the above, it eventually became evident through empirical studies of grading practices that K-12 report card grades are multidimensional measures of a variety of factors, both cognitive (those related to student achievement and understanding of the material) and non-cognitive (those which are not directly tied to student achievement), that assess, as well as motivate, student learning based on what teachers value in student work. These factors often include such things as achievement, substantive engagement, persistence, improvement, and even the consequences of grades on students' success and feelings about their own competence [13]. These studies align well with teachers' perceptions of their own grading practices as determined by surveys and interviews where teachers brought up the inclusion of non-cognitive factors in the grades they assign, along with many teachers expressing a desire to grade fairly, which to them meant using multiple sources, incorporating effort, and making grading policies clear to students. Context and professional judgment is sometimes included as well, rather than relying solely on an impersonal grading algorithm. However, teachers' beliefs and values determine the purpose and extent of the impact that non-cognitive factors have on the grades they assign, and these vary between teachers on both an individual and group level, sometimes between different teachers in the same school and sometimes even between students who have the same teacher (due to differing contexts) [13].

On the group level, modern elementary school teachers largely think of grades as being more about communicating information to students and parents while secondary school teachers think of them more in terms of classroom management (accounting for student behavior and completion of work), along with placing a higher value on exams [13]. While it is likely that there exists a range of grading schemes in higher education, especially when grading standards are considered to be a matter of academic freedom in the U.S., one might suspect that these trends continue and that college instructors incorporate a variety of factors into the grades they assign, but emphasize the exam-based achievement side of grades more than K-12 teachers do. Once again, it would be beneficial for additional studies to be conducted so a more definitive determination can be made, but in the author's experience this could certainly be argued in the case of many introductory college physics courses, including those involved in this study where, as discussed in "Site, Sample, and Population," timed exams and quizzes are the primary basis for course grades while participation and homework completion (non-cognitive factors) play a significant, but rather limited, role. It could also be helpful for more thorough evaluations of grades in K-12 schooling, as well as higher education, to be done using formal factor analysis (both exploratory and confirmatory based on previous research) in order to better understand and account for the different components that go into student grades, both in general and in particular situations.

One possible way to address discrepancies between the relative weights given to different components of grades (exam scores, engagement, completion of work, etc.) by different teachers would be to standardize these weights by requiring the proportion of grades attributed to each component be the same (or at least within some predetermined range) across all students, or at least all students within the same grade level and type of course. This may not go over well with many teachers though and academic freedom, combined with grading autonomy, would make it difficult to enforce. A related but distinct possibility would be to give separate scores for different components of grades, as is the case with standards-based grading (a form of grading that describes where a student is relative to predetermined standards for different aspects of a given level of a given subject) [13]. This would recognize the importance of different student attributes but would systematically and consistently distinguish between them, and it would even be possible to include a section on context or professional judgement. In some ways, this would actually be analogous to how certain grading metrics, like Grade-Point-Averages (GPAs), are aggregated across different classes, but grades in specific classes are still typically reported as well, since this would simply be a further disaggregation within classes. However, these sorts of suggestions ignore the fact that some teachers consider behavior that promotes academic achievement to be part of academic achievement and therefore, may discount any attempts to separate such things on principle [13]. Furthermore, even though differential impact of a test on different groups is not considered "bias" on its own in measurement theory (as discussed in more detail below), many teachers are rightfully concerned about the negative material consequences that low grades, as well as their intersection with other student characteristics (like race, gender, and socioeconomic status), can have under the current system. Such concerns may cause teachers to be reluctant to assign separate grades for different student attributes, knowing that low scores in certain areas will likely reproduce various forms of institutional violence (like forcing these students' future-selves into low-wage jobs or inadequate housing) even when reported in conjunction with high scores in other areas [16, p. 478]. The only way to remedy this concern would be to dismantle capitalism and other forms of oppression by restructuring the way that society functions such that people are no longer punished simply for having low academic achievement scores.

One last point that should be brought up in this discussion of grades' criterion-related validity is that in K-12 schooling, grades are known to predict drop-out rates and other measures of success and failure in subsequent levels of schooling [13]. However, one could argue that this is circular reasoning since it is unsurprising that measures of "success" in schooling relate to other measures of "success" in schooling regardless of what underlying construct(s) any of these (generalized) tests actually measure. It is interesting to note, however, that standardized achievement tests do not predict such things nearly as well, which raises further questions about the meaning, interpretation, and uses of both standardized achievement tests and grades [13].

Finally, it is important to consider the content validity of grades' achievement component. Content validity refers to the degree of alignment between a test and the content (both knowledge, and reasoning processes and skills) that those who take the test have been taught [16, p. 257-258], with one way of determining content validity being an evaluation of the test by a content expert. More specifically in the context of this study, since overall Lecture grades for the courses involved here are primarily based on quizzes and exams, it could be argued that they are largely reflective of student achievement. However, even if this is true in some sense, it still would not necessarily guarantee that these grades accurately measure the same types of material (both knowledge, and ways of solving problems, which is really emphasized in these courses) that is taught to the students who take these courses. Grades in these courses reflecting some sort of physics achievement, even if true, also would not necessarily guarantee that they measure the types of achievement that physics instructors believe they do, or that application reviewers for jobs, internships, graduate school, etc. believe they do, which is not necessarily the same as what physics instructors believe they do (thereby adding another confounding layer to questions of content validity). There are a few reasons for these potential discrepancies. One reason is that, not only are quizzes and exams just one possible form of assessing achievement among many, but the quizzes and exams used in the courses that were part of this study also tend to be timed and necessitate short, written responses (as is typical in introductory physics courses). Considering how students' understanding of physics principles is the construct of interest for this study, as well as the construct of interest for Lecture instructors and, presumably, for application reviewers who take grades from these courses into account, this format has a negative impact on student grades relative to the construct of interest by both introducing a degree of construct-irrelevant variance (variation in test scores that is due to factors other than the construct of interest) and causing some degree of construct underrepresentation (when a test does not fully incorporate all parts of the construct of interest, or weighs different parts of this construct disproportionately compared to their relevance to the construct) [16, p. 261].

A few possible sources of construct-irrelevant variance in the courses that were part of this study are speediness, the ability to perform under pressure, reading comprehension, writing abilities, and handwriting. These factors lead to some variance in quiz and exam scores that is independent of the construct of interest (understanding of physics principles) and may even introduce bias into these scores, as well as the Lecture grades that are heavily dependent on them. Bias in measurement theory is defined as a situation where a test yields systematically different scores for respondents from different groups who are actually at the same level on the underlying construct (referred to as their "true score" in Classical Test Theory) [16, p. 279-280]. Therefore, the factors mentioned above could produce bias in the courses that were part of this study because of the different ways that these issues affect different groups [16, p. 479]. For instance, if physics quizzes and exams require a certain level of reading comprehension or writing ability, then scores on them are likely to be lower for non-native English speakers than for their native English-speaking counterparts because of language barriers that have nothing to do with physics knowledge or understanding. Similarly, timed exams are known to lower scores for women and girls disproportionately more than they lower scores for men and boys. Much, though not necessarily all, bias of this sort, as well as construct-irrelevant variance more broadly, could be adequately addressed by quizzes and exams that follow principles of universal design, which seeks to eliminate construct-irrelevant factors (like wordy language, speediness, and cultural references) while also specifically emphasizing those factors which may systematically disadvantage members of certain groups (including by offering multiple examination formats) [16, p. 503].

Construct underrepresentation in these courses could come from restricting the types of achievement that grades in these courses capture due to quizzes and exams focusing almost exclusively on an individual's direct knowledge and calculational abilities while largely leaving out such things as lab skills, oral processing, and the ability to productively expand upon others' ideas. This too could potentially bias course grades because of differences in what different groups are socialized to value, such that certain groups may be better at the types of skills that are being often tested in these courses while others may be better at those that are not. It is also quite possible that students in these courses end up focusing on the types of behaviors and skills that will yield the highest possible grades (given the primary use of grades as ranking and sorting mechanisms, as described previously), in which case this construct-irrelevant variance and construct underrepresentation could easily have the effect of shaping what students prioritize when trying to learn the material. It would therefore be beneficial if future research could determine the types of skills that are measured by conventional physics exams and quizzes (perhaps using factor analysis), along with how this impacts students' study habits. It would then be up to the physics education community to decide if these are desirable skills to measure, as well as whether they are the only skills that should be measured, and to adjust accordingly.

Going beyond testing mechanisms and confounding factors, it is often not even the case that all aspects of the material which is covered in introductory physics courses, such as those involved in this study, ends up being represented on the quizzes and exams that any given student takes, and even when it comes to the material that is represented, the breakdown of how much each topic contributes to students' grades can vary widely. It would be helpful to both students and application reviewers if physics instructors could come to some sort of consensus on which topics should be covered in introductory courses, their breakdown in terms of grading, and the breakdown of difficulty levels within each topic, and then largely stick to this structure. However, since this is unlikely to occur, each individual instructor should at least take the time to think through their own feelings on these matters and be explicit with themselves, their peers, and their students about such things through personal specification tables which list the content they believe is important (and will be covering) for a given course and the level that they will be covering each topic at (in that course) [16, p. 269].

A solution to the need for content experts to evaluate quiz and exam problems is less clear since it would not be reasonable to require every quiz or exam problem to go through this process and most physics instructors would consider themselves to be content experts when it comes to introductory physics principles (and really, any physics content that they are teaching in most cases) anyway. There is also some dispute over who would qualify as an expert in this regard and whether, for instance, professional physicists who are not teaching faculty would be effective at making these sorts of determinations [17]. However, there might be some need for the courses involved in this study to have their quizzes and exams, or at least a range of sample problems for instructors to base quizzes and exams off of, go through some sort of expert evaluation since these courses are not only taught with a different format than traditional lecture-based physics courses, but they also use a different curriculum (and thus, textbook and set of homework problems) that physics instructors who have been trained through more conventional means may not be familiar with. Obviously, none of these suggestions can be enforced is most cases because of academic freedom and grading autonomy, but individual instructors are still able to adhere to best practices when creating assessments (like quizzes and exams) and should be encouraged to do so.

One final issue around the content validity of the courses that were part of this study is the potential mismatch between what is taught during DL and what appears on quizzes and exams since, as discussed previously, DL is taught by a TA while quizzes and exams for these courses are usually designed and administered by the Lecture instructor and given during lecture sections. The lectures for these courses are supposed to coincide with what is taught in DL, but this is not always the case, both in terms of the standard DL curriculum (which is rather rigid) and how DL is actually taught in practice (which can vary between DLs based on TA style, potential TA modifications, etc.). This issue involves an extremely relevant question for this study since the goal of this study is to evaluate the relationship between Lecture grades and classroom characteristics, especially DL size, which requires Lecture grades to be a relatively strong measure of what is taught in DL. For the purposes of this study, it was assumed that despite all of their flaws, overall Lecture grades in the courses involved here at least measure the same content as what is taught in DL, but it might be worth conducting a rigorous analysis of this relationship in the future. Similarly, it was assumed that the students who take the courses that were part of this study are malleable and that their physics learning depends on the classroom environment and instructional guidance, since if this is not the case then DL and Lecture characteristics would clearly and fundamentally have no impact on their understanding of physics. This too is a question that could use further study.

Reliability

While the validity of grades is important, both in general and as a framework to help guide the conclusions and limitations of this study, evaluating the merits of grades does not end there. Beyond the extent to which intended uses and interpretations of grades match what they actually measure, another concern that anyone who uses grades to judge student performance (from researchers to instructors to application reviewers) should have is simply the degree to which grades (as a test in the generalized measurement theory sense) consistently measure something, be it academic achievement or a broader multidimensional mix of traits (i.e. regardless of what it is the test is actually measuring). This is where reliability comes in [16, p. 121].

In general, it is possible that approaches to grading which separate out different components of grades, like standards-based grading, could make grades more reliable because each grade would then measure a single construct. On top of this, even under more traditional models of grading, it has been found that as assignments are aggregated up to class (or Lecture) grades, the reliability of grades tends to increase regardless of what it is that they are measuring [13]. This means that in a typical scenario, a greater proportion of students' "true scores" on whatever construct(s) a given class' grade is measuring is represented by their overall grade in that course than by their grade on any particular assignment, which makes sense because one would expect the effect of random error to decrease as more assignments are taken into account [16, p. 161-162]. This implies that overall class grades are likely to be a better proxy for students' understanding of physics concepts in introductory physics courses (including those involved here) than grades on individual assessments would, assuming that both overall class grades and individual assessment grades overwhelmingly measure physics understanding rather than a multitude of constructs (and for the courses involved in this study, it is likely that overall Lecture grades and individual quiz and exam grades do primarily measure some sort of physics understanding, as has already been discussed). It is true, however, that even in these courses, overall Lecture grades likely measure other constructs to some degree, whereas quiz and exam grades likely measure other factors to a lesser extent. Because of this, it is still possible that grades on certain individual assessments might be better indicators of students' physics understanding than overall Lecture grades and this may be a good question for future research. It is for this reason, as well as the fact that overall Lecture grades for the courses involved in this study are largely just aggregated grades from individual quizzes and exams, that for the purposes of this study, the reliability of individual assessments is a vital part of any discussion around the reliability of overall Lecture grades, and so the reliability of individual assessments is where this section now turns.

For any given individual assessment, there are a variety of reliability concerns to consider. First off, there are questions of internal reliability (whether or not different items are measuring the same construct(s)) and alternate forms reliability (whether or not different forms of the assessment are measuring the same construct(s)) if different forms of the assessment are given. These two types of reliability are difficult to study in general since they depend on the nature of the assignment in question, which can vary tremendously across instructors, classes, schools, etc. because of the large degree of autonomy that teachers usually have in coming up with assignments, both in K-12 schooling and especially in university settings where this is often a matter of academic freedom. However, it is once again the case that individual instructors, including those who teach the courses involved in this study, could adhere to best practices and should strive to do so. More specifically, even if it is not required that they do so, Lecture instructors could potentially address these concerns by conducting their own reliability studies by, for instance, finding coefficient alpha (which is essentially the correlation between different items or problems on a given test) [16, p. 137-139] for their own quizzes and exams, or finding the correlation between alternate forms of a typical quiz or exam [16, p. 195-197] (though these analyses may be rather cumbersome, especially for a Lecture instructor with a lot of other tasks to complete). Furthermore, administrators should encourage them to do so, both informally and through direct incentives like including these practices in an instructor's (properly compensated) workload and taking these practices into account when conducting teacher evaluations and making decisions about promotions.

A similar line of reasoning can be applied to quiz or exam item bias (using the same definition of bias discussed above, except for individual items within a test as opposed to a test as a whole) where it is difficult to study item bias in general because any such bias would be unique to individual problems written by individual instructors for individual classes (though one could conduct general studies of item bias on common types of problems, which is certainly something that exists in introductory physics courses). However, as with internal and alternate forms reliability, whenever possible it would be good practice for individual instructors to conduct their own analyses of item bias and to discard items accordingly, and they should be encouraged to do so [16, p. 483-499].

Regardless of any suggestions for future practices though, there is no way to tell what the internal or alternate forms reliability of the quizzes and exams that were used by the Lectures involved in this study during the period of study were, nor is there any way to tell how biased they were as a result of either individual item bias or more holistic factors like those described earlier. And yet, such things certainly affect the reliability and validity of Lecture grades and thus, the accuracy and meaning of regression analyses that include Lecture grades as a variable (especially the outcome variable that is being used as a proxy for student understanding of underlying concepts), so these unknown pieces of information present a definite limitation to this study that should not be ignored.

Another important aspect of reliability to consider is inter-rater reliability, or the consistency of grades that different graders would assign to a particular individual assessment (for instance, student 12's quiz number 3 or maybe even problem 2 of student 12's quiz number 3) [16, p. 210-212]. This aspect of reliability is possible to study in general and many such studies (focusing on K-12 teachers, though their conclusions likely extend to university instructors as well given the individualized nature of grading preferences) have been conducted, particularly during the early 1900s [13]. These studies largely showed a significant degree of variation between different teachers (who were also graders) of about 5 points on a 100 point scale, though a few studies disagreed with this conclusion [13]. The primary sources of variation were an inability to distinguish between assignments of similar "merit" (which can be conceptualized as random error), differences between teachers' grading standards, and differences between the relative weights that teachers assigned to different aspects of an assignment [13]. It is not much of a stretch to imagine that bias could be a relevant factor here as well (even if it is not one that many prominent academics thought about during the early 20th century) since grader biases have the potential to show up in the grades that they assign, and different graders have different types and levels of bias. This variability in the grades that different teachers who were involved in these studies would give to the same assignment eventually led to the development and implementation of letter grading in an attempt to reduce the effect of rater uncertainty on grades, which bolsters the argument put forward in "Levels and Outcome Variable" that letter grades (or their numerical equivalent) is an appropriate outcome variable to use in this study [13].

However, the methodologies used in these early studies of inter-rater reliability had their flaws. For example, teachers were often sent assignments to grade without specific grading criteria [13]. Because of this, some of the uncertainties in grading that were identified by these studies could be reduced through a range of improvements, from using better grading criteria that incorporates student input to more collaboration among teachers when it comes to grading practices and standards [13]. Wider adoption of standards-based grading as discussed previously could also help by parsing out the different components of grades and effectively standardizing the aforementioned weights [13].

Another practice that would help accomplish these goals would be the formation and implementation of better grading criteria in the form of more rigorous and standardized rubrics [18] and/or grading by category (where similar mistakes are grouped together into a category such that grades are effectively determined by one or more categories, which may or may not be mutually exclusive) [19]. There is already some evidence to suggest that categorical grading is more consistent (reliable) across different graders than more traditional grading methods, both for quantitative (calculations and diagrams) problems and especially for conceptual ones [9]. While part of this may be due to the 4.5-point grade scale (which directly corresponds to letter grades) having fewer possible grades than the 10-point scale (which directly corresponds to percentage grades), it turns out that graders using the 10-point scale typically assign almost the same number of unique grades as graders using the 4.5-point scale [14]. Research also suggests that inter-rater reliability, both between different graders who are all using categorical grading, as well as between categorical grading and more traditional grading methods, is higher for quantitative questions than it is for conceptual ones [9]. However, as noted above, this research also suggests that conceptual questions are often better at gauging physics understanding (i.e. are more valid for this construct) than quantitative questions, so both of these factors should be accounted for when creating physics assessments [9]. In general, grade determination is most difficult, and grader consistency is lowest, in situations that fall on the lower end of the grading scale but still involve substantial student work (as opposed to being blank or close to it) [9]. Additional research would need to be done before making stronger assertions about the reliability of these techniques or their application to the courses that were part of this study, but they seem to have at least some promise of generalizability. Either way though, the courses in this study already use rubrics and grade by category and each problem is usually graded by a single TA anyway, so while there is always room for improvement, inter-rater reliability is probably not a major limitation of this study.

One last way to potentially increase the inter-rater reliability of individual assessments, at least in physics courses, would be to use assessments that require less subjective grading by, for example, employing multiple choice questions that approximate certain aspects of free-response questions. While the two will never be equivalent and many physics instructors are skeptical of multiple choice physics problems, there are some preliminary results suggesting that it is possible for multiple choice questions to mimic their free-response counterparts under the right circumstances [20]. This is especially true if incorrect answers on the multiple choice version of a problem conform to common mistakes that students often make in the free-response version (the fact that common mistakes can be categorized in this way is also the basis for grading by category), and provided that different levels of partial credit are given to incorrect multiple choice answers in a similar manner to how partial credit would be assigned to similarly incorrect free-response answers [20]. This discussion leads into a much longer discussion about the relative merits of different assessment formats, but purely from the perspective of measurement theory, it would require, at minimum, much more extensive research on alternate forms reliability between free response problems (and assessments, like quizzes or exams, which are composed of several problems) and their multiple choice counterparts by developing a large question bank which includes both versions of each question; administering them to a large, representative sample; and finding the correlation in scores between the two versions. Assessments formed from these questions would also have to be evaluated for internal reliability and free response problems would have to be checked for inter-rater reliability during the research portion of such a program in order to make sure that all of these items are measuring the same construct and that when the multiple choice version of a question is compared to the free response version, there is a well-defined and agreed upon free response (partial credit) score to use as a reference point. Finally, the validity of such assessments would need to be studied to make sure they are fully evaluating all aspects of the desired underlying construct(s) (i.e. all aspects of physics knowledge and understanding that they are meant to evaluate).

Taken together, all of the above implies that overall class grades are a reflection of a range of important traits that the application reviewers who most frequently use them are likely to be interested in (provided these reviewers actually care about evaluating applicable constructs rather than simply reproducing social hierarchies, intentionally or not), from academic achievement to communication skills to team work to effort and perseverance. However, the degree to which different characteristics contribute to grades can differ quite substantially across different classes, instructors, graders, schools, etc. and a lot of people, likely including many application reviewers, do not realize this and instead believe that grades are purely a reflection of academic achievement, which at the moment is both not typically the case and also not necessarily desirable given the importance of various other student attributes. Furthermore, it must be acknowledged that even the achievement component of grades does not always reflect the full range of academic ability that one might expect them to and plenty of people are probably unaware of this as well. False beliefs about the nature and interpretation of grades can therefore impact the decisions that are made based on them in a way that does not properly reflect their true meaning or appropriate uses, which is something that should be addressed by administrators, managers, politicians, and others who have power over relevant policies under the current system.

In some cases, though, grades do largely reflect academic ability and not much else, and it would seem as if overall Lecture grades for the courses involved in this study are among these cases, meaning these grades are a fairly good proxy for physics understanding (whether or not the types of physics understanding that they reflect correspond to the types of material that is taught in DL or the content that instructors and others believe they reflect, which are separate questions) and are therefore a fairly good outcome variable to use in this study. These Grades are clearly not perfect, though, and there are still some problems with using them in this way that future research will hopefully shed more light on. Perhaps it will turn out that grades on certain individual assessments, like final exams, are better for this purpose, or that something entirely different from grades, like scores on the FCI or an analogous assessment, would be best, but in the meantime, overall Lecture Grades appear to be a relatively good approximation of student understanding.

Appendix C: Sequential HLM Model Equations in This Study

For all of the models used in this study, i is an index labeling individual students (or really, observations, but this distinction has little practical meaning), j is an index labeling DLs, and k is an index labeling Lectures.

Also note that these equations are written entirely in a level-by-level form, as opposed to a composite form, and refer to analyses involving data from the regular academic year (a few modifications would be needed to get the equations for analyses involving data from the summer, as noted and briefly discussed below).

Null Model

Level 1:

$$\mathrm{Grade}_{ijk} = \beta_{0jk} + \varepsilon_{ijk}$$

Level 2:

$$\beta_{0jk} = \gamma_{00k} + u_{0jk}$$

Level 3:

$$\gamma_{00k} = \pi_{000} + \nu_{00k}$$

Individual Model

Level 1:

$$\mathrm{Grade}_{ijk} = \beta_{0jk} + \beta_{1jk} * \mathrm{Male}_{ijk} + \beta_{2jk} * \mathrm{UnS}_{ijk} + \sum_{m=3}^{16} \beta_{mjk} * \mathrm{RaceandEthnicitym}_{ijk}$$

$$+ \sum_{m=17}^{22} \beta_{mjk} * \mathrm{USCitizenshipStatusm}_{ijk} + \beta_{23jk} * \mathrm{Grad}_{ijk} + \beta_{24jk} * \mathrm{Repeat}_{ijk}$$

$$+ \beta_{25jk} * \mathrm{LecStart}_{ijk} + \beta_{26jk} * \mathrm{GPA}_{ijk} + \beta_{27jk} * \mathrm{Units}_{ijk} + \varepsilon_{ijk}$$

Level 2:

$$\beta_{0jk} = \gamma_{00k} + u_{0jk}$$

$$\beta_{mjk} = \gamma_{m0k} \ \ \forall m \in \{1, \ldots, 27\}$$

Level 3:

$$\gamma_{00k} = \pi_{000} + \nu_{00k}$$

$$\gamma_{m0k} = \pi_{m00} \ \ \forall m \in \{1, \ldots, 27\}$$

Where $\mathrm{RaceandEthnicity3} = \mathrm{AF}, \mathrm{RaceandEthnicity4} = \mathrm{AI}, \ldots, \mathrm{RaceandEthnicity16} = \mathrm{UnE}$, following the order they appear in “Level Choices and Predictor Variables” while skipping WH since this is the reference category for Race and Ethnicity in this study.

Similarly, $\mathrm{USCitizenshipStatus17} = \mathrm{PR}$, $\mathrm{USCitizenshipStatus18} = \mathrm{NI}, \ldots, \mathrm{USCitizenshipStatus22} = \mathrm{UnC}$, following the order they appear in “Level Choices and Predictor Variables” while skipping Cit since this is the reference category for U.S. Citizenship Status in this study.

Note that LecStart was included here because it is a level 1 predictor variable during the regular academic year, but during the summer it would not appear in this way and the last few terms in the Level 1 equation above would instead be:

$\beta_{25jk} * \mathrm{GPA}_{ijk} + \beta_{26jk} * \mathrm{Units}_{ijk} + \varepsilon_{ijk}$

On a related note, during the summer $m \in \{1, \ldots, 26\}$

DL Model

Level 1:

$$\mathrm{Grade}_{ijk} = \beta_{0jk} + \beta_{1jk} * \mathrm{Male}_{ijk} + \beta_{2jk} * \mathrm{UnS}_{ijk} + \sum_{m=3}^{16} \beta_{mjk} * \mathrm{RaceandEthnicitym}_{ijk}$$

$$+ \sum_{m=17}^{22} \beta_{mjk} * \mathrm{USCitizenshipStatusm}_{ijk} + \beta_{23jk} * \mathrm{Grad}_{ijk} + \beta_{24jk} * \mathrm{Repeat}_{ijk}$$

$$+ \beta_{25jk} * \mathrm{LecStart}_{ijk} + \beta_{26jk} * \mathrm{GPA}_{ijk} + \beta_{27jk} * \mathrm{Units}_{ijk} + \varepsilon_{ijk}$$

Level 2:

$$\beta_{0jk} = \gamma_{00k} + \sum_{n=1}^{6} \gamma_{0nk} * \mathrm{DLSizen}_{jk} + \sum_{n=7}^{14} \gamma_{0nk} * \mathrm{DLTimen}_{jk} + \gamma_{015k} * \mathrm{ROS}_{jk} + \gamma_{023k}$$

$$* \mathrm{Mean\_GPA}_{jk} + \gamma_{022k} * \mathrm{Mean\_Units}_{jk} + \gamma_{020k} * \mathrm{Mean\_Male}_{jk} + \gamma_{021k}$$

$$* \mathrm{Mean\_LecStart}_{jk} + u_{0jk}$$

$\beta_{mjk} = \gamma_{m0k}\ \ \forall m \in \{1, \ldots, 27\}$

Level 3:

$\gamma_{00k} = \pi_{000} + \nu_{00k}$

$\gamma_{0nk} = \pi_{0n0}\ \ \forall n \in \{1, \ldots, 23\}$

$\gamma_{m0k} = \pi_{m00}\ \ \forall m \in \{1, \ldots, 27\}$

Where $\mathrm{DLSize1} = \mathrm{RlySm}, \mathrm{DLSize2} = \mathrm{Sm}, \ldots, \mathrm{DLSize6} = \mathrm{RlyLg}$, following the order they appear in “Level Choices and Predictor Variables” while skipping Stand since this is the reference category for DL sizes in this study.

Similarly, during the regular academic year $\mathrm{DLTime7} = \mathrm{DL8}$,

$\mathrm{DLTime8 = DL105, \ldots, DLTime14 = DL1542}$, following the order they appear in “Level Choices and Predictor Variables” while skipping DL1417 since this is the reference category for DL start times during the regular academic year in this study.

During the summer, the sum on $\mathrm{DLTimen_{jk}}$ would only go to n = 12 and the corresponding DL Time variables would be $\mathrm{DLTime7 = DL95, DLTime8 = DL11, \ldots, DLTime12 = DL1717}$, following the order they appear in “Level Choices and Predictor Variables” and Table 7 in “Analysis Format and Summary Data” while skipping DL1217 since this is the reference category for DL start times during the summer in this study. The coefficients in front of the terms that come after the DLTime terms would also be renumbered accordingly.

On a related note, during the summer n ∊ $\{1, \ldots, 21\}$. Also, note that the notes about LecStart and m above apply here as well.

Final Model

Level 1:

$$\begin{aligned}\mathrm{Grade_{ijk}} &= \beta_{0jk} + \beta_{1jk} * \mathrm{Male_{ijk}} + \beta_{2jk} * \mathrm{UnS_{ijk}} + \sum_{m=3}^{16} \beta_{mjk} * \mathrm{RaceandEthnicitym_{ijk}} \\ &\quad + \sum_{m=17}^{22} \beta_{mjk} * \mathrm{USCitizenshipStatusm_{ijk}} + \beta_{23jk} * \mathrm{Grad_{ijk}} + \beta_{24jk} * \mathrm{Repeat_{ijk}} \\ &\quad + \beta_{25jk} * \mathrm{LecStart_{ijk}} + \beta_{26jk} * \mathrm{GPA_{ijk}} + \beta_{27jk} * \mathrm{Units_{ijk}} + \varepsilon_{ijk}\end{aligned}$$

Level 2:

$$\begin{aligned}\beta_{0jk} = \gamma_{00k} + \sum_{n=1}^{6} \gamma_{0nk} * \mathrm{DLSizen}_{jk} + \sum_{n=7}^{14} \gamma_{0nk} * \mathrm{DLTimen}_{jk} + \gamma_{015k} * \mathrm{ROS}_{jk} + \gamma_{023k} \\ * \mathrm{Mean\_GPA}_{jk} + \gamma_{022k} * \mathrm{Mean\_Units}_{jk} + \gamma_{020k} * \mathrm{Mean\_Male}_{jk} + \gamma_{021k} \\ * \mathrm{Mean\_LecStart}_{jk} + u_{0jk}\end{aligned}$$

$$\beta_{mjk} = \gamma_{m0k} \;\; \forall m \in \{1, \ldots, 27\}$$

Level 3:

$$\gamma_{00k} = \pi_{000} + \pi_{001} * \mathrm{LecSize}_{k} + \sum_{p=2}^{3} \pi_{00p} * \mathrm{Termp}_{k} + \sum_{q=4}^{42} \pi_{00q} * \mathrm{Instructorq}_{k} + \nu_{00k}$$

$$\gamma_{0nk} = \pi_{0n0} \;\; \forall n \in \{1, \ldots, 23\}$$

$$\gamma_{m0k} = \pi_{m00} \;\; \forall m \in \{1, \ldots, 27\}$$

Where $\mathrm{Instructor4} = \mathrm{Ins1}, \mathrm{Instructor5} = \mathrm{Ins2}, \ldots, \mathrm{Instructor42} = \mathrm{Ins40}$, following a numerical order while skipping the Lecture instructor reference category for each analysis (which are listed at the end of “Regression Results and Variance Discussion”).

During the regular academic year, Term2 = Winter for 7A but Fall for 7B and 7C while Term3 = Spring for 7A and 7B but Winter for 7C. Note that predictor variables related to academic term do not appear in analyses involving data from the summer and thus, in such analyses the Instructorq variables (and corresponding coefficients) would be renumbered accordingly.

Also note that the above notes about LecStart, m, DL start times, and n apply here as well.

Lastly, note that the slope coefficients in this Final Model that are associated with the DLSizen categorical predictor variables (π_{0n0} for $n \in \{1, \ldots, 6\}$), as well as (though to a lesser degree) the slope coefficient that is associated with the LecSize continuous predictor variable (π_{001}), are the primary focus of this study and will be used to help answer the main research question.