Title

Quantifying Multilevel Effects in the LOCUS Statistics Assessment: An Analysis

ISBN

9798247993063

Author

Divecha, Anirudh

Publication Date

2026-06-09

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
Los Angeles

Quantifying Multilevel Effects in the LOCUS Statistics Assessment: An Analysis

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Statistics

by

Anirudh Divecha

2026

ABSTRACT OF THE THESIS

Quantifying Multilevel Effects in the LOCUS Statistics Assessment: An Analysis

by

Anirudh Divecha
Master of Science in Statistics
University of California, Los Angeles, 2026
Professor Robert Gould, Co-Chair
Professor Hongquan Xu, Co-Chair

The Levels of Conceptual Understanding in Statistics (LOCUS) assessment is a standardized test designed to measure conceptual understanding in statistics education. Educational assessment data, such as LOCUS, usually exhibit hierarchical structure, with students nested within teachers, schools, and state. Ignoring this structure can often lead to misidentifying important sources of variation and can lead to misleading interpretations of student performance. This thesis applies hierarchical linear modeling (HLM) and other statistical analyses to the LOCUS dataset in order to examine and quantify teacher variation in exam performance.

The thesis begins with an exploratory analysis of the LOCUS dataset. Then, hierarchical linear models are fit separately across ten LOCUS exam forms to determine within-teacher and between-teacher variance components. Null models indicate substantial clustering at the teacher level across all forms. Full models incorporate fixed effects for school year, population of test takers, testing phase, and state in order to evaluate how much between-teacher variation can be explained by observable covariates. Results show that these covariates

do account for a substantial portion of teacher-level variance, with state contributing the largest reductions in variance. However, estimated state effects are unstable across forms. After adjustment for observable covariates, residual teacher-level ICCs remain consistent across forms, generally ranging from approximately 0.15 to 0.20.

This thesis provides evidence that variation in LOCUS performance does exist at the teacher level, even after accounting for fixed effects. These findings contribute to a better understanding of the factors associated with students' conceptual understanding of statistics over the years of LOCUS administration.

The thesis of Anirudh Divecha is approved.

Drago Plecko

Hongquan Xu, Committee Co-Chair

Robert Gould, Committee Co-Chair

University of California, Los Angeles

2026

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

Analyzing student performance in educational settings and exams plays an important role in understanding learning outcomes for educators and schools and can significantly inform policy decisions and teaching methods. In many cases, analyses often ignore sources of variability that exist due to the nature of educational systems and datasets. For example, students tend to be naturally nested within teachers, schools, districts, and states which creates hierarchical dependencies that must be taken into account in these kinds of statistical analyses. Ignoring this hierarchical structure can often lead to incorrect conclusions about where variation in student performance originates [GH07].

The LOCUS (Levels of Conceptual Understanding in Statistics) exam was used to study these structural dynamics in the specific context of statistics education [WFJ15]. The data that was compiled from this assessment includes student performance across multiple exam forms, academic years, populations of test takers, teacher information, and state. The LOCUS dataset reflects educational settings where differences in teacher, curriculum, and student populations can contribute to exam performance. The hierarchical nature of the dataset also makes it well suited for hierarchical modeling approaches so that variation at different levels of the educational system can be analyzed.

In this thesis, we employ hierarchical linear modeling (HLM) and other assessment focused analysis techniques to the LOCUS dataset in order to understand student performance [RB02]. We begin by confirming the comparability of exam forms through analyzing the equating set – which are shared questions that are designed to anchor performance across

forms. Building on previous LOCUS exam developer's work, we verify the validity of the LOCUS exam on the full dataset which contains more observations, states, and exam years [JdH23]. We then go on to evaluate how much variation in exam scores can be attributed to teachers and how much of this variation can be explained by covariates such as academic school year, population of test takers, and when during the academic cycle the test was given. We begin by fitting null models to evaluate whether there is any substantial clustering at the teacher level and then extend these models by incorporating fixed effects. We first focus on a subset of the data that is restricted to the state of California and then go on to analyze the full dataset which contains multiple states. Throughout the analysis, our aim is to quantify the variation associated with teachers rather than to estimate causal teacher effects.

The contributions of this thesis include a variance decomposition across all ten LOCUS forms in order to quantify how much score variation is associated with teachers before and after accounting for fixed effects. We also show that while state explains a large and consistent share of between teacher variation, the estimated state effects are unstable across forms. This potentially indicates that state in the models acts as a statistical absorber of local LOCUS participation patterns rather than an indicator of state educational systems. Lastly, we provide evidence that the magnitude of teacher level variation depends on the population of test takers and curricular context of administration. Overall, by comparing models across multiple exams and evaluating changes in variance components, this thesis provides insights into the factors associated with variation in LOCUS exam performance and, more broadly, students' conceptual understanding of statistics.

Research Questions

This thesis is guided by the following research questions. Are the LOCUS exam forms comparable enough using the equating set? How much of the variation in LOCUS exam performance is attributable to differences between teachers? To what extent do covariates

explain the between-teacher variation? Which covariates account for the largest reductions in teacher-level variance, and how stable are those effects across exam forms? After accounting for covariates, how much variation remains and how should it be interpreted? Does teacher-level variation differ by testing population or curricular context?

CHAPTER 2

Data

2.1 LOCUS Dataset Overview

The LOCUS dataset consists of student response/answers to multiple standardized statistics assessments designed to evaluate conceptual understanding of statistics. The data has been collected over several years, from around 2014–2025, and there are ten exam forms that can be analyzed, each with varying difficulty levels and numbers of questions ranging from 23 to 50 questions. All together there are five exams each with a Form 1 and Form 2 version for a total of ten exams. A summary of the LOCUS exam data can be seen in the Table 2.1. Each observation in the dataset corresponds to a student's attempt on the exam and contains their response to each question among other information such as student demographics, teacher, and school.

2.2 Structure of the Data

The dataset is organized as an excel file with multiple sheets where each sheet corresponds to a specific exam form (as noted and labeled underneath "Exam Title" in Table 2.1). Within each sheet, rows represent individual students and columns include various data information and question level response information. The initial columns contain demographic variables (gender, race, primary language) and school information (school name, school town, school state, school zip, school country). Additional variables include the type of student taking the exam, name of the teacher administering the exam, and timestamps indicating when the

Table 2.1: Summary of LOCUS exam forms, difficulty levels, and number of questions

Exam Title	**Exam Difficulty**	**Number of Questions**
BegIntForm1_23	Beginner/Intermediate	23
BegIntForm2_23	Beginner/Intermediate	23
BegIntForm1_30	Beginner/Intermediate	30
BegIntForm2_30	Beginner/Intermediate	30
IntAdvForm1_23	Intermediate/Advanced	23
IntAdvForm2_23	Intermediate/Advanced	23
IntAdvForm1_30	Intermediate/Advanced	30
IntAdvForm2_30	Intermediate/Advanced	30
BegIntAdvForm1_50	Beginner/Intermediate/Advanced	50
BegIntAdvForm2_50	Beginner/Intermediate/Advanced	50

exam was completed. Subsequent columns indicate question level responses corresponding to the question number on the exam and student's recorded answer to the question. For example, a column could be labeled "Question 1" and a row will contain whether the student answered 1, 2, 3, or 4 for their answer choice. This structure allows for analysis of the entire exam but also allows for fine grain analysis at the question level.

2.3 Data Cleaning and Key Features

For each exam sheet, the first 13 columns were preserved as categorical variables while the remaining question response columns (Question 1, Question 2, ... Question n) were coerced to numeric format. Across all exam forms, no student row observations were removed and all exam sheets were verified to have successfully loaded. There was also a companion answer key file with sheets that corresponded to the exam form. The answer key data file contained question level information such as question number, question title, content

Table 2.2: Description of key variables in the LOCUS dataset

Variable Name	**Category**	**Type**	**Description**
Form	Exam Info	Categorical	Identifier for exam form or version
Gender	Demographic	Categorical	Student gender
Test Taker Hispanic	Demographic	Categorical	Hispanic/Latino identification
Test Taker Race	Demographic	Categorical	Student race category
Test Taker Language	Demographic	Categorical	Primary language of student
School Name	School	Categorical	Name of the school
School Town	School	Categorical	City of the school
School State	School	Categorical	State of the school
School Country	School	Categorical	Country of the school
Pop. of Test Takers	Demographic	Categorical	Type of student population (AP student, middle school student, college student)
Teacher Name	Instructional	Categorical	Teacher identifier
Test Finished	Temporal	Datetime	Timestamp when exam was taken
School Year	Temporal/Derived	Categorical	School Year from timestamp
Testing Phase	Derived	Categorical	early vs. late testing administration within academic cycle
Q1, Q2, . . . , Qn	Item Response	Numeric	Raw student response to each question
Q1_correct, Qn_correct	Item Response	Binary	Indicator for correct or incorrect response
exam_scores_raw	Outcome/Derived	Numeric	Total number of correct responses
exam_scores_percent	Outcome/Derived	Numeric	Percentage score on exam

category, difficulty level, and correct answer. A scoring algorithm was built to score each of the exams leading to new features being derived, which are shown in Table 2.2, and are labeled as derived in the "Category" column. The scoring algorithm generated correctness indicators (0 or 1) for every question, treated missing responses as incorrect, and added two derived summary variables to each student record: raw score (exam_scores_raw) and percent score (exam_scores_percent). Raw score is the total number of correct responses and percent score is the raw score divided by the total number of questions on an exam form. By cleaning up the data in this way, analysis can be conducted on the equating set of questions such as percent scores on the equating set which then is used to assess comparability between different versions of the exams.

In addition, some variables were derived from the exam completion timestamp ("Test Finished"). The variable "School Year" is derived as a typical academic school year which spans from August to July. Thus, tests that were administered during their respective school year were labeled and treated as categorical. As a concrete example, an exam completed in September 2019 and one completed in May 2020 were both classified as belonging to the 2019–2020 school year. Another variable, "Testing Phase", was derived in order to capture when within the academic cycle an exam was administered and was categorized into two levels: "early" and "late". This is because we suspect that students taking the exam earlier in the academic cycle are most likely "new" to statistics, while those students who take it later might have had greater exposure to statistical concepts and instruction. Thus, to account for potential differences associated with timing of administration and exposure to statistical instruction, the "Testing Phase" variable was created. Because academic cycles differ between high school and college level test takers, the binning for early or late was defined separately for school populations. Since AP Statistics, Grades 10–12, and Grades 6–9 can test at any point during a full academic school year, exams completed between August-December were categorized as "early" and exams taken between January-July were considered "late". For the Introductory Statistics (college) population, labeling was done

differently because colleges often follow a semester system. Thus, exams completed during August–October and January–February were labeled as "early" since these would be considered early in the semester while the remaining months were considered "late".

Overall, the key features which play an important role in this analysis include exam form, academic year the test was administered, and population of test takers. The main outcome variable being studied was student performance on each exam which is given by their percent score on the exam.

2.4 Data Limitations and Considerations

There are several limitations of the data that should be acknowledged. These limitations include imbalanced group sizes across teachers and schools, sparse observations for certain key features, and overall variability in sample sizes across exam forms. In terms of sparse observations, demographic features, such as gender or test taker race, were not collected after the year 2018 for all LOCUS exams. Thus, instead of subsetting the data and losing observations, these features were mostly not used in models due to their sparsity. In addition, in the context of HLM modeling, not all of the exam forms contained the hierarchical depth to support multilevel clusters which led to specific considerations when specifying models. These limitations are addressed throughout the analysis to make sure statistical conclusions are valid and interpretable.

CHAPTER 3

Preliminary Analyses

Before conducting the primary hierarchical linear modeling analysis, some preliminary analyses were conducted in order to understand the LOCUS data and confirm the validity of the exam itself. The next few sections include an overview of the exploratory data analysis, comparisons of question level difficulty across forms, and evaluation of the equating set.

3.1 Exploratory Data Analysis

We first started by obtaining a descriptive overview of the LOCUS dataset such as who took the exam, the score distributions across forms, and what data quality issues exist, if any. This EDA informed much of the subsetting of the data for modeling decisions.

3.1.1 Sample Composition and Population Exclusions

Table 3.1 reports the number of test takers per exam form. The dataset spans over 70,000 test administrations across the ten exam forms. Form 1 tended to have more test administrations than Form 2 reflecting how schools/teachers utilized the forms over academic years.

Table 3.2 shows that there are seven distinct groups of test takers across exam forms. Note that in this particular case, "Introductory Statistics" refers to college/university level introductory statistics course students while "Grades 10-12" refers to high school level students. Mean score differences across groups vary pretty drastically: AP Statistics students score on average 10-20 percentage points higher than Grades 10-12 Non-AP students, and

Table 3.1: Number of Test Takers by Exam Form

Exam	**Number of Students**
BegInt Form 1 (23)	15,390
BegInt Form 2 (23)	12,592
BegInt Form 1 (30)	3,952
BegInt Form 2 (30)	1,881
IntAdv Form 1 (23)	16,972
IntAdv Form 2 (23)	5,410
IntAdv Form 1 (30)	7,287
IntAdv Form 2 (30)	3,162
BegIntAdv Form 1 (50)	2,390
BegIntAdv Form 2 (50)	1,203

as expected, teachers who took the exam score higher.

Since teachers, health professionals, and examinees coded as "Other" are not the focus of this study, they are excluded from the modeling that follow. This exclusion is consistent with the LOCUS assessment's use as a diagnostic tool for students enrolled in statistics courses. It is also worth noting that the data contains a small number of international countries, primarily from Tianjin (China), Ankara (Turkey), and several other locations outside the United States. Students from these countries are also not included in the modeling.

In addition, Table 3.3 shows the counts for the four retained student populations across the exam forms and reveals a few structural features of the data. BegInt forms mostly consist of Grades 10–12 while more advanced forms tend to be dominated by AP Statistics and Introductory Statistics. This alignment between population ability and form level difficulty shows that the forms were administered to students based on the student's expected statistical ability. In addition, several forms are sparsely populated or empty, specifically the Grades 6-9 column. What this means is that the population fixed effect in the full HLM

Table 3.2: Population of Test Takers Aggregated Across All Exam Forms

Population	**Count**	**Mean Score Range (%)**
Grades 10–12, Non-AP	21,049	∼38–51
AP Statistics	19,844	∼60–68
Introductory Statistics	16,767	∼42–60
Other	6,774	∼46–73
Teachers	4,096	∼62–79
Grades 6–9	1,671	∼42–60
Health Professionals	38	∼61–79

Note. Mean score range reflects variation across all ten exam forms within each population group.

models in subsequent analyses operate over a different set of population groups which is worth keeping in mind when comparing variance reductions across forms. This is because with more heterogeneous populations for an exam, there is more room for teacher variance to be absorbed from the population covariate in contrast to forms that have a more homogeneous population mix. Thus, Table 3.3 is not only descriptive but is connected to the explanatory power of the population covariate in the HLM models that follow.

3.1.2 Score Distributions

Table 3.4 shows the score distributions for students across all ten exam forms. As testing difficulty increases, mean and median scores tend to increase as well. Form 2 scores are consistently higher than Form 1 scores across all exams. After consulting with LOCUS developers, it was confirmed that in some but not all instructional cases, Form 1 was used as pre-test and Form 2 was used as a post-test. This suggests that higher Form 2 scores may simply reflect increased exposure to statistical content over the academic cycle rather than differences in form difficulty. Because the LOCUS dataset does not contain unique student identifiers, Form 1 and Form 2 scores cannot be linked at the student level, making it difficult

Table 3.3: Population Counts by Form

Form	**AP Statistics**	**Grades 10–12 (Non-AP)**	**Grades 6–9**	**Introductory Statistics**
BegInt Form 1 (23)	1,411	8,935	526	1,282
BegInt Form 2 (23)	1,199	7,056	727	1,204
BegInt Form 1 (30)	1,088	997	263	942
BegInt Form 2 (30)	124	733	98	494
IntAdv Form 1 (23)	6,332	1,932	0	7,226
IntAdv Form 2 (23)	1,097	278	0	3,061
IntAdv Form 1 (30)	4,746	910	22	717
IntAdv Form 2 (30)	1,989	157	0	684
BegIntAdv Form 1 (50)	1,268	17	0	603
BegIntAdv Form 2 (50)	373	0	0	498

Note. Excludes Teachers, Health Professionals, and Other populations.

to confirm this pre-test/post-test interpretation directly. However, labeling when during the academic cycle an exam was administered provides a partial way to account for this and is the direct motivation for deriving the "Testing Phase" variable (which was introduced in Section 2.3). By classifying the time frame within the academic cycle for when each testing administration occurred, subsequent models can condition on whether students were likely at the beginning or end of their statistics instruction without having to link student IDs across forms.

In addition, comparing the 23 versus 30 question forms within each difficulty level shows that adding seven additional questions raises the median score slightly but does not meaningfully increase score dispersion which is measured by the IQR. Overall, this suggests that the additional questions raise overall scores but do not substantially improve discrimination between student ability levels.

Table 3.4: Score Distribution Summary Statistics by Exam Form (Student Examinees Only)

Exam Form	**Mean**	**Median**	**Q1**	**Q3**	**SD**	**IQR**
BegInt Form 1 (23)	41.4	39.1	30.4	52.2	17.3	21.7
BegInt Form 2 (23)	46.4	43.5	30.4	60.9	20.0	30.4
BegInt Form 1 (30)	49.7	46.7	33.3	66.7	20.8	33.3
BegInt Form 2 (30)	49.3	50.0	36.7	60.0	17.2	23.3
IntAdv Form 1 (23)	50.8	52.2	34.8	65.2	19.1	30.4
IntAdv Form 2 (23)	56.1	56.5	43.5	69.6	19.2	26.1
IntAdv Form 1 (30)	54.9	56.7	40.0	70.0	18.8	30.0
IntAdv Form 2 (30)	55.9	60.0	40.0	73.3	20.8	33.3
BegIntAdv Form 1 (50)	59.1	60.0	48.0	72.0	17.5	24.0
BegIntAdv Form 2 (50)	63.8	66.0	54.0	76.0	17.0	22.0

Note. All scores and summary statistics are reported as percentages. Non-student populations are excluded.

3.1.3 Missingness Rates

Figure 3.1 displays question level missingness rates for all exams. Missingness in this context reflects questions a student left unanswered or skipped. Overall, missingness remains pretty low and a consistent pattern that emerges is that missingness rates tend to gradually increase towards the end of the exam. This trend can be attributed to testing fatigue or time constraints near the end of the assessment rather than issues with questions themselves. Because missingness rates remain low and systematic across forms, they are unlikely to meaningfully distort score distributions or equating analyses.

3.1.4 Geographic Coverage

Table 3.5 shows the distribution of teachers and students across the states represented in the LOCUS dataset with all ten exam data pooled together. Much of the data is heavily concentrated in a small number of states with California and Arizona having the higher number of observations. The "Forms Present" column indicates the number of distinct LOCUS exam forms (out of 10) in which an administration of an exam happened in that state. The "Teachers" column represents the distinct number of teachers from that state.

In addition, Appendix A reports the full state level distribution of teachers, schools, and students for each of the ten LOCUS exam forms. These tables provide the evidence underlying the cluster structure claims and modeling decisions made in later sections. Some specific claims include that the teacher to school ratio is approximately 1:1 across forms for many of the states which motivates a two level model specification. In addition, the tables in Appendix A show that the majority of states contribute only one or two teachers on any given form which motivates treating state as a fixed rather than random effect.

The dominance of California in both teacher and student counts further motivated the California only analysis as it is the only state that has a significant number of observations.

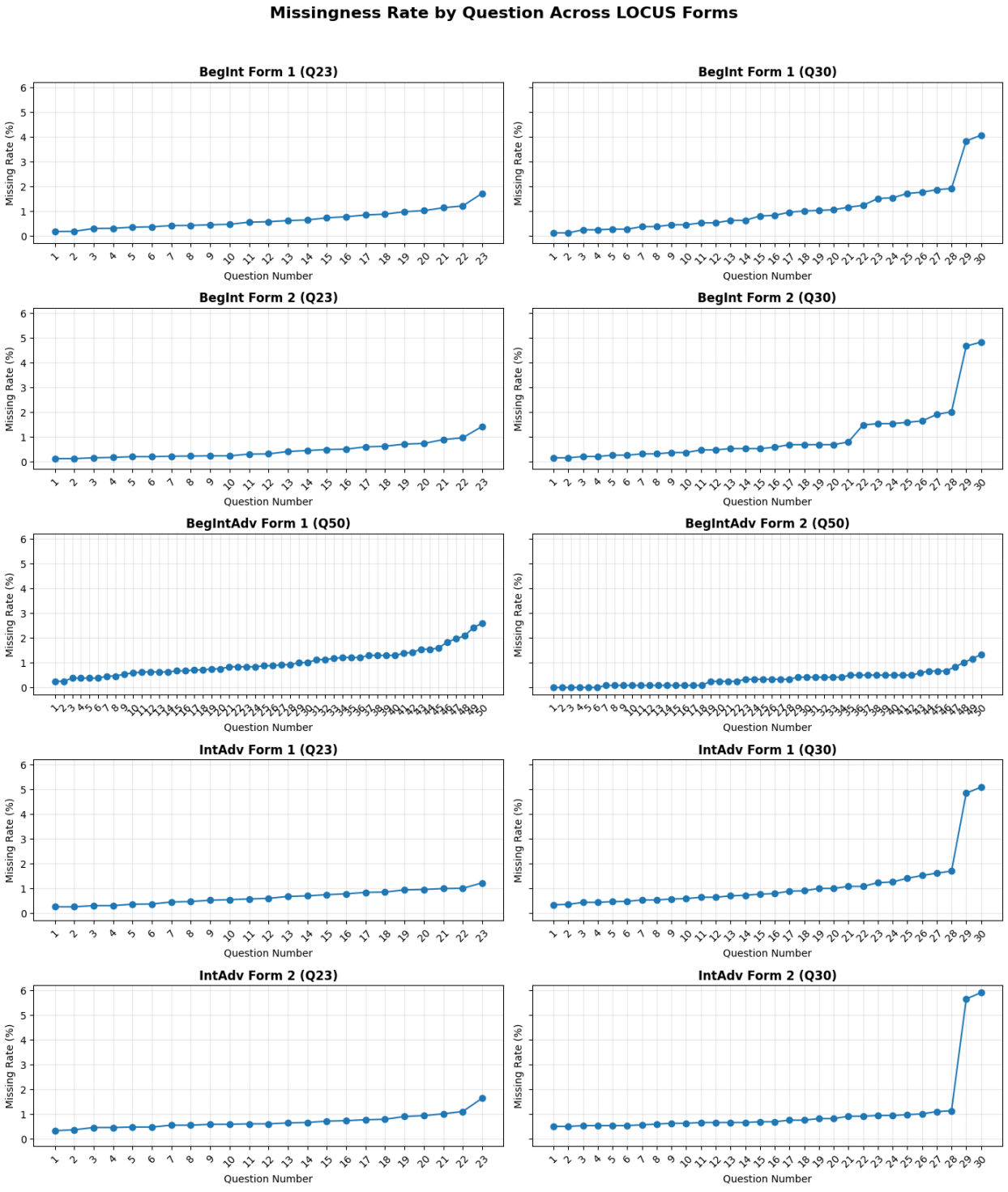

Figure 3.1: Missingness rate by question across LOCUS exam forms. Each panel displays the percentage of questions skipped for that particular question within a given form.

Table 3.5: Distribution of Teachers and Students by State for All LOCUS Exams

State	Teachers	Students	Forms Present
California	53	18,937	10
Florida	20	4,943	10
Virginia	15	3,460	9
New York	13	2,996	8
Georgia	10	615	6
Wisconsin	10	1,777	9
Arizona	9	11,955	8
Massachusetts	9	450	8
Ohio	9	1,672	8
New Jersey	7	1,149	8
All other states ($n = 23$)	71	11,035	NA

3.1.5 Score Distributions Over Time

Figure 3.2 shows mean exam scores over time for LOCUS exams. Sample sizes (n) are labeled at each data point and do vary across years.

Across all forms, mean scores seem to be within approximately 35–70%, with most forms clustering between 45% and 60%. Note that the BegIntAdv Q50 form has the most variability from year to year which can mostly be attributed to the smaller and more variable sample sizes. The spike to nearly 90% in 2019 for Form 2 (n = 2) and the sharp drop in 2018 (n = 41) for the BegIntAdv Q50 exam shows that these mean estimates should be interpreted with caution at small n.

Overall, the year over year mean score data support the claim that the LOCUS exam has produced relatively stable mean performance for more than a decade, while also revealing instability in forms with smaller annual samples.

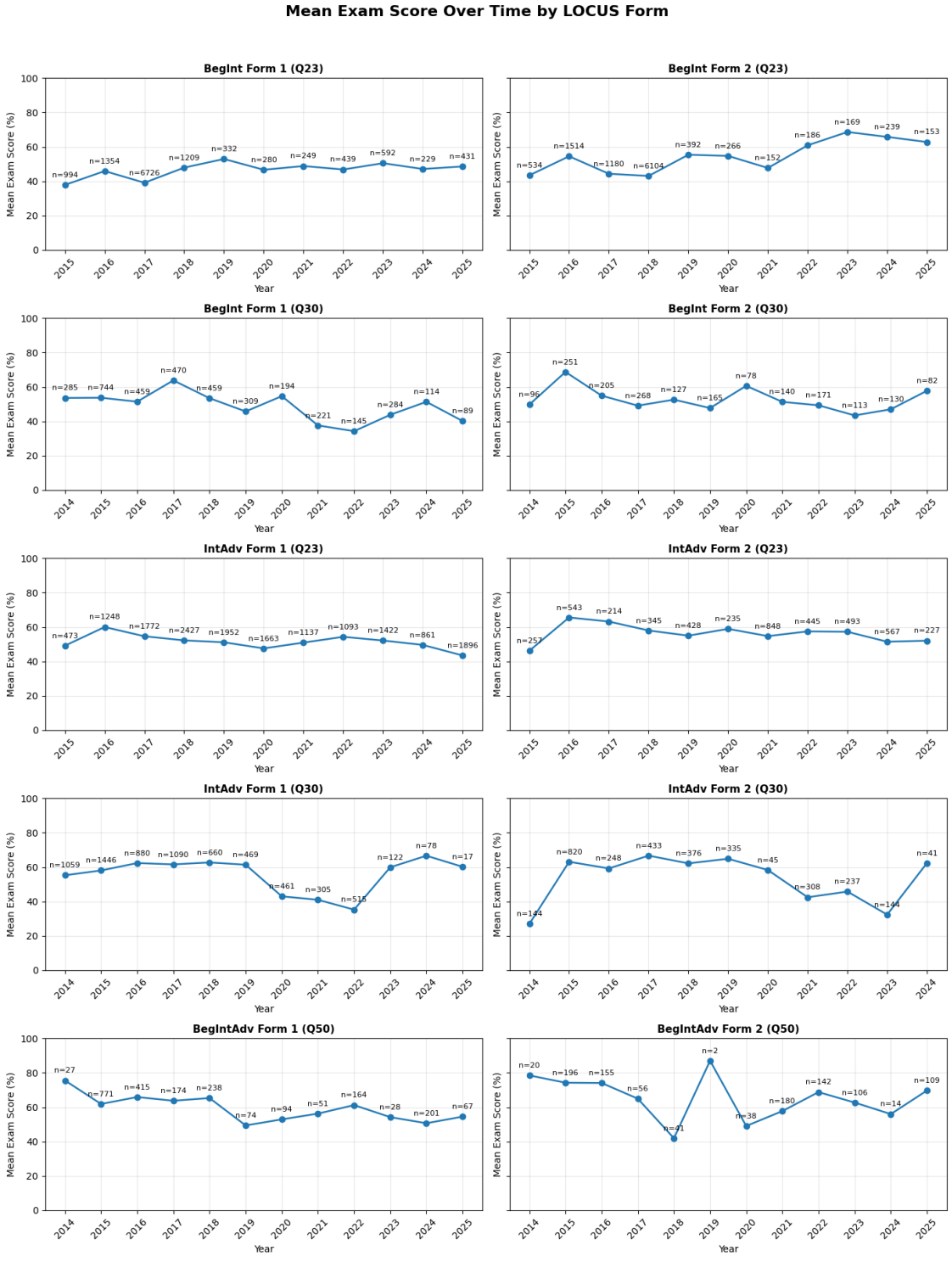

Figure 3.2: Mean LOCUS exam scores over time by form

3.2 Equating Set Analysis

All LOCUS exams have a common set of ten anchor questions, also known as the equating set. The equating set consists of questions Q2, Q4, Q6, Q8, Q10, Q12, Q15, Q18, Q20, and Q22 which are completely identical across all exams; this means that any student who took the LOCUS exam answered the same ten questions from the equating set. The validity of the LOCUS assessment was previously evaluated by LOCUS exam developers through question analysis and calibration on a small sample of students across six states [JdH23]. The analysis presented in this thesis confirms that the properties previously established still hold on the full LOCUS dataset, which is substantially larger and spans more states and years. Similarly to the original validation, we evaluate whether performance on the equating set reflects overall testing ability and whether it generalizes to questions that are not part of the equating set.

Usually for the equating set to be a good set, it must satisfy two conditions. First, performance on the equating set should reflect overall testing ability [Ang84]. What that means is that students who perform well on the equating items should be strong students overall. Second, performance on the equating set should generalize beyond the equating set questions themselves. What this means is that students who score well on the equating set questions should also perform well on the non-equating set of questions on the exam. Usually, if the second condition holds, the equating set can be considered a reliable metric for measuring general ability and used for cross form score comparison [Liv04].

In order to evaluate these two conditions, Pearson correlations were computed between each test takers' equating set score and two other measures: (1) their total exam score (including equating items), and (2) their score computed only on non-equating questions.

3.2.1 Descriptive Statistics For Equating Set

We begin by reporting summary statistics for scores on only the equating set in Table 3.6. Scores are expressed as percentages of the ten equating items answered correctly.

Table 3.6: Equating Set Score Summary Statistics by Exam Form

Exam Form	**Mean (%)**	**Median (%)**	**Q1 (%)**	**Q3 (%)**
BegInt Form 1 (23)	39.7	40.0	30.0	50.0
BegInt Form 2 (23)	42.8	40.0	30.0	60.0
BegInt Form 1 (30)	46.9	40.0	30.0	70.0
BegInt Form 2 (30)	43.5	40.0	30.0	60.0
IntAdv Form 1 (23)	55.3	60.0	40.0	70.0
IntAdv Form 2 (23)	61.5	60.0	50.0	80.0
IntAdv Form 1 (30)	60.8	60.0	50.0	80.0
IntAdv Form 2 (30)	61.9	70.0	40.0	80.0
BegIntAdv Form 1 (50)	61.8	60.0	50.0	80.0
BegIntAdv Form 2 (50)	66.0	70.0	50.0	80.0

Note. Scores are reported as percentages of equating items answered correctly.

Overall, equating set score distributions vary across exam difficulty levels. Beginner-Intermediate students score lower on average which is consistent with the fact that students in this group would generally have lower overall statistics proficiency. Intermediate-Advanced and Beginner-Intermediate-Advanced examinees score considerably higher due to having better statistical proficiency. In all cases, the interquartile range spans 30–40 percentage points, indicating similar spread in equating set performance with differences in medians. These properties are essential for the correlation analysis described below.

3.2.2 Pearson Correlation Analysis

Table 3.7 shows Pearson correlations between scores on the equating set and two other performance outcomes: total score on the exam and score on questions that were not in the equating set.

Table 3.7: Pearson correlations between equating set scores and exam outcomes

Exam	ρ **(Equator vs. Total)**	ρ **(Equator vs. Non-Equator)**
BegInt Form 1 (23)	0.854	0.563
BegInt Form 2 (23)	0.859	0.612
BegInt Form 1 (30)	0.880	0.726
BegInt Form 2 (30)	0.815	0.578
IntAdv Form 1 (23)	0.892	0.638
IntAdv Form 2 (23)	0.885	0.642
IntAdv Form 1 (30)	0.869	0.687
IntAdv Form 2 (30)	0.888	0.744
BegIntAdv Form 1 (50)	0.812	0.699
BegIntAdv Form 2 (50)	0.779	0.661

Correlations between equating set scores and total exam scores were high across all ten forms, ranging from $\rho = 0.75$ to $\rho = 0.89$. However, it's important to note that these correlations are somewhat inflated because the equating items are a subset of the total score. Overall, the more informative analysis is the correlation with the non-equating exam scores.

Correlations between equating set scores and non-equating exam scores were moderate to strong across all forms, ranging from $\rho = 0.50$ to $\rho = 0.74$. The lack of weak or close to zero correlations with non-equating scores indicates that the equating set functions similarly across the Beginner-Intermediate, Intermediate-Advanced, and Beginner-Intermediate-Advanced exams. In the original validation study, the equating set was designed to have

difficulty and content distribution similar to the overall assessment and the correlations shown here on the full dataset confirm the design choices that were made by the LOCUS test creators [JdH23]. The equating set can therefore be considered a representative subset of test content: students who score well on the equating set tend to score well on the non-anchor portion of the exam as well.

3.2.3 Equating Set Difficulty Across Forms

The difficulty of question j is defined as the proportion of students who answered the questions correctly:

$$p_j = \frac{\text{Number of students who answered question } j \text{ correctly}}{\text{Total number of students}}$$

Values of p_j closer to 1 indicate easier questions, while values closer to 0 indicate more difficult questions.

Figure 3.3 plots the proportion correct on each equating question for Form 1 against the corresponding proportion correct for Form 2. The diagonal dashed line represents equal difficulty so points lying near this line mean that the question functioned with comparable difficulty across forms. Across all five panels in Figure 3.3, the equating questions hover closely around the diagonal indicating stable equating. Once again, the stability across paired forms shown here provide further confirmation of the findings outlined in the validation study but on a larger dataset [JdH23].

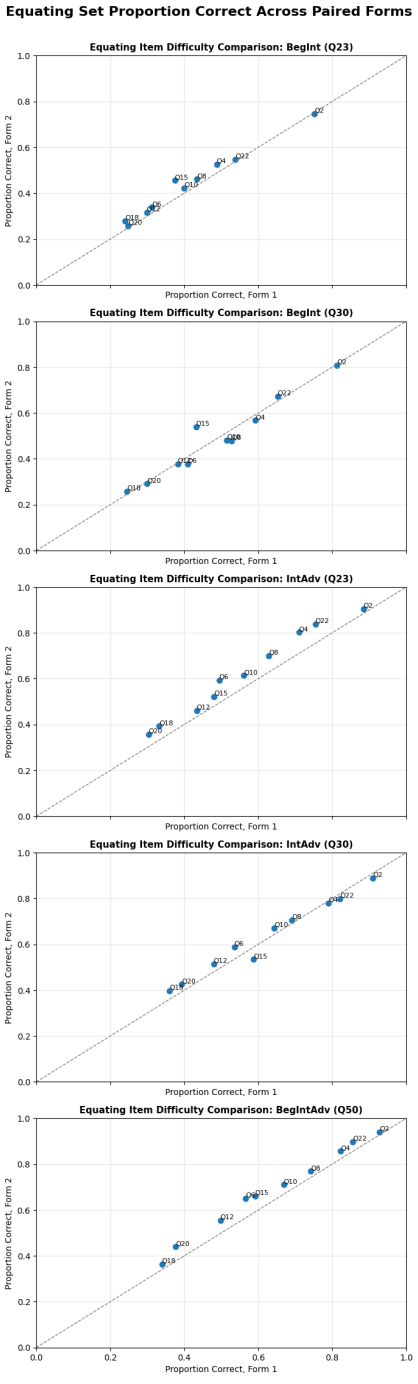

Figure 3.3: Equating set proportion correct across paired forms

3.3 Summary

All together, the preliminary analysis sets up the hierarchical linear modeling that follows. Exam score distributions revealed a consistent pattern of Form 2 scores exceeding Form 1 scores, attributable in many cases to a pre-test/post-test administration rather than differences in form difficulty. The equating set analysis confirms the design rationale that was proposed in the original LOCUS validation paper [JdH23]. Pearson correlations between equating set performance and non-equating exam performance ranging from 0.56 to 0.74 across all forms, indicates that students who performed well on the anchor items also tended to perform well on the overall exam. These findings support the use of the equating set as a good comparability metric across forms. With a clear understanding of the data's structure and limitations, the next chapter turns to the hierarchical linear modeling framework used to partition variance in student performance across students and instructors.

CHAPTER 4

Methodology

4.1 Motivation for Using HLM

The LOCUS data has a clear hierarchical structure: individual students are nested within teachers, teachers are nested within schools, and schools are nested within states. This structure leads to observations violating independence assumptions thus requiring a multi-level modeling framework.

Traditional regression approaches assume that observations are independent. In the LOCUS dataset, this is violated because students who are taught by the same teacher are likely to have similar teaching experiences and/or classroom environments. As a result, student exam scores may be correlated to students within the same teacher versus students who were taught by different teachers. Ignoring this clustering can lead to misleading results which leads us to model the data using hierarchical linear modeling techniques [GH07].

Hierarchical linear modeling (HLM), also known as multilevel modeling, is a modeling technique that accounts for the nested structure by allowing model parameters to vary across clustering units such as teachers. Variation in student performance can be broken down into within-teacher and between-teacher components allowing for a better understanding of the sources of variability.

In the context of the LOCUS dataset, using HLM enables us to address key research questions: How much of the variation in exam performance is attributable to differences between teachers? To what extent do observable factors, such as academic year or student

population, explain these differences?

4.2 Initial HLM Data Structure

After the data preparation, the outcome variable that was selected was the student's percent score on the exam (exam_scores_percent). In this data, teacher names serve as the main clustering level in all models reported.

Certain features of the data also influenced certain modeling choices which are discussed here. Although teachers are nested under schools, the empirical teacher-to-school ratio is close to 1:1 in most exam forms as noted in the tables in Appendix A. This means that school-level variance and teacher-level variance are one and the same. Any attempts to fit three level models (students nested in teachers nested in schools) usually produced singular fits or convergence warnings due to this 1:1 ratio. In addition, when state was then considered as a level-3 grouping variable, only some states (notably California and Arizona) contributed enough teachers to have stable estimation; the majority of states contributed only one or two teachers, which once again provides essentially no information for estimating variance between states [LH21].

Given these constraints, all models reported in this thesis use a two-level random intercept structure with students nested within teachers, and treat state as a fixed-effect covariate. The decision to omit school and state as random levels is purely driven by data limitations and not by the reasoning that those levels do not matter.

4.3 Model Specification

For the analysis, two models are fit for each exam form: a null (unconditional) model that has only the teacher random intercept, and a full (conditional) model that adds fixed effects. Comparing the null and full model allows for the analysis of the extent to which teacher-level

variance can be explained by the measured covariates.

4.3.1 Null Model

The null model is specified as:

$$Y_{ij} = \beta_0 + u_j + \epsilon_{ij},$$

where Y_{ij} denotes the exam score for student i taught by teacher j, β_0 represents the overall mean exam score, $u_j \sim \mathcal{N}(0, \sigma^2_{\text{teacher}})$ is the random effect associated with teacher j, and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2_{\text{student}})$ is the student-level residual error. This allows the intercept to vary across teachers.

4.3.2 Full Model

We can extend the null model to a full model by including relevant covariates as fixed effects. For example, we included exam year, population of test takers, and testing phase as predictors to account for some of the variation in exam scores across time and student groups. Testing phase is included to control for whether students were administered the exam early or late in the academic cycle. The resulting model is given by:

$$Y_{ij} = \beta_0 + \beta_1(\text{School Year}_j) + \beta_2(\text{Population}_j) + \beta_3(\text{Testing Phase}_j) + u_j + \epsilon_{ij}.$$

In this model, β_1, β_2, and β_3 represent the fixed effects of exam year, population of test takers, and testing phase respectively, while the random intercept u_j captures teacher level deviations from the overall mean. By comparing this model to the null model, we can assess whether these covariates explain variation in student outcomes and reduce between-teacher variation.

All models were estimated using the `lme4` package in R using the `lmer` function [BMB15].

This is a standard package in HLM estimation and provides an efficient framework for estimating multilevel model coefficients and variance components.

4.4 Variance Decomposition and ICC

In order to quantify the degree of clustering in the data, we extract the variance components from the hierarchical models and compute the intraclass correlation coefficient (ICC) [RB02, HH07]. The teacher ICC is defined as the proportion of total variance in the score that is attributable to differences between teachers:

$$\text{ICC} = \frac{\sigma^2_{\text{teacher}}}{\sigma^2_{\text{teacher}} + \sigma^2_{\text{student}}}$$

A higher ICC indicates that a large share of the variation in student performance is associated with teacher to teacher differences. A lower ICC suggests that most of the variability occurs within teachers at the student level rather than between teachers.

To quantify how much of the variation between teachers is explained by the inclusion of covariates, we compute the proportional reduction in variance (PRV) which is defined as

$$\text{PRV}_{\text{teacher}} = \frac{\sigma^2_{\text{teacher, null}} - \sigma^2_{\text{teacher, full}}}{\sigma^2_{\text{teacher, null}}}$$

[RB02].

4.5 Modeling Strategy

The modeling framework is applied separately to each of the ten exam forms because exam forms vary in difficulty and in the population of students who take them. Thus, pooling forms simply does not make sense when modeling exam outcomes. In addition, fitting models to each form separately allows us to assess whether the ICC patterns and the explanatory power of the covariates are consistent across forms. The thought is that replication across forms

strengthens conclusions, while differences between form results flags these forms as needing additional scrutiny.

The two analyses that follow in the next chapter apply this framework to two facets of the data. The first analysis is done to only students in California, which served as an initial scoping exercise to confirm that HLM is appropriate before extending to the full dataset. The second analysis is then done on the full dataset, in which state is included as a fixed effect and the larger sample size supports more stable estimation.

CHAPTER 5

Results

5.1 California-Restricted Analysis

California is a natural starting point for a few reasons: it contributes the largest share of test administrations in the LOCUS dataset and it is the home state of UCLA and therefore of interest for this thesis.

5.1.1 Setup

The full model for the California analysis is described here:

$$Y_{ij} = \beta_0 + \beta_1(\text{School Year}_j) + \beta_2(\text{Population}_j) + \beta_3(\text{Testing Phase}_j) + u_j + \epsilon_{ij}.$$

Models are fit separately for each exam form and for forms in which the sample size or cluster structure does not support stable estimation the corresponding variance components are reported as NA in the results table (Table 5.1).

5.1.2 Results

Table 5.1 summarizes the estimated variance components and ICC at the teacher level for both the null and the full model. The column "PRV" represents the percentage reduction in estimated teacher variance (column 4 to column 5) and is the equivalent of the PRV that was defined previously in Section 4.4. This metric quantifies the extent to which observed teacher

differences can be explained by these additional covariates in the full model. The column "ΔAIC" reports the change in Akaike Information Criterion between the reduced and null models [Aka74]. Negative ΔAIC values indicate that the full model fits the data better than the null model, accounting for the additional parameters introduced by the covariates.

Table 5.1: Teacher-Level Variance Decomposition by Exam Form (California Sample)

Exam	**ICC (Null)**	**ICC (Full)**	**Var (Null)**	**Var (Full)**	**PRV**	**ΔAIC**
BegInt Form 1 (23)	0.351	0.114	73.558	17.379	76.4%	-27.349
BegInt Form 2 (23)	0.328	0.338	98.463	100.178	-1.7%	-30.638
BegInt Form 1 (30)	0.539	0.371	191.032	83.624	56.2%	-154.885
BegInt Form 2 (30)	NA	NA	NA	NA	NA	NA
IntAdv Form 1 (23)	0.526	0.219	284.771	69.337	75.7%	-256.097
IntAdv Form 2 (23)	0.331	0.383	144.065	174.956	-21.4%	-62.072
IntAdv Form 1 (30)	0.476	0.281	185.342	63.605	65.7%	-410.463
IntAdv Form 2 (30)	0.482	0.597	200.651	262.218	-30.7%	-153.659
BegIntAdv Form 1 (50)	0.329	NA	108.565	NA	NA	NA
BegIntAdv Form 2 (50)	0.256	NA	66.000	NA	NA	NA

Note. ICC = intraclass correlation coefficient. PRV = percent reduction in teacher-level variance relative to the null model. ΔAIC = full model AIC minus null model AIC.

5.1.3 Variance Decomposition and ICC

Across exam forms, the null model ICC values range from 0.256 to 0.539, which are much higher than the 0.15 to 0.25 range that is recommended as typical for school level ICCs of academic achievement [HH07]. These ICC values suggest strong clustering at the teacher level and provide justification for the use of a hierarchical modeling framework.

In addition, the inclusion of covariates (year, population, and testing phase) in the full model leads to changes in teacher variance across all exams. In some cases, we observe

substantial reductions in teacher variance. For example, in BegInt Form 1 (23), teacher variance decreases from 73.558 to 17.379, corresponding to a 76.4% reduction. Similarly, BegInt Form 1 (30) exhibits a 56.2% reduction, and IntAdv Form 1 (23) shows a 75.7% reduction. For these forms, the covariates explain a large share of between-teacher variability.

However, this pattern is not consistent across forms as three forms show small negative reductions such as BegInt Form 2 (23) at -1.7%, IntAdv Form 2 (23) at -21.4%, and IntAdv Form 2 (30) at -30.7%. Negative percentages indicate that estimated teacher variance rose after adding the covariates and can either reflect estimation instability or that these covariates redistribute instead of absorb variance for these specific forms.

Note that BegInt Form 2 (30), BegIntAdv Form 1 (50), and BegIntAdv Form 2 (50) contain NA values in the table. This is due to model convergence issues and instead of reporting biased or incorrect estimates, we simply exclude them from the table and analysis.

5.1.4 Model Fit and Comparison

Model fit, as measured by changes in AIC, generally improves with the inclusion of covariates, as indicated by negative ΔAIC values across the forms that could be estimated. The improvement is large for several forms such as IntAdv Form 1 (30) with ΔAIC = -410.463 and IntAdv Form 1 (23) with ΔAIC = -256.097. These results suggest that the full models provide a better fit to the data in most cases.

5.1.5 Discussion

Overall, high null-model ICCs across the LOCUS exams indicate strong clustering at the teacher level. This implies that teacher variation plays a significant role in the distribution of student performance across the exams. In addition, the variability in results across exam forms highlights the importance of analyzing each exam separately, as the influence of covariates and the magnitude of teacher effects are not uniform across assessments. Several

exam forms could not be fit reliably due to small cluster sizes, motivating using the entire dataset for a larger sample size. The next section extends the same modeling framework to the full multi-state dataset, where the inclusion of state as a fixed effect can potentially capture a source of between-teacher variation.

5.2 Full-Dataset Analysis

5.2.1 Setup

The full model for the entire dataset analysis is described here:

$$Y_{ij} = \beta_0 + \beta_1(\text{School Year}_j) + \beta_2(\text{Population}_j) + \beta_3(\text{Testing Phase}_j) + \beta_4(\text{State}_j) + u_j + \epsilon_{ij}.$$

Models were attempted that nested teachers within schools but in all cases, fits were singular or did not converge, making the three-level specification not feasible. All results reported below therefore use the two-level (students within teachers) model specification and instead add state as a fixed-effect instead of as a random effect.

5.2.2 Results

Table 5.2 reports the variance components, ICCs, percent reduction in teacher level variance, and ΔAIC for each of the ten exam forms. With the inclusion of school year, population of test takers, testing phase, and state as fixed effects, all ten forms produced models that converged in both the null and full cases with non-zero estimates of teacher variance. This is a meaningful improvement over the California analysis, where small sample sizes led to some forms not able to obtain a fit; the larger samples in the full dataset support stable estimation on every form.

Table 5.2: Teacher-Level Variance Decomposition by Exam Form (Entire Dataset)

Exam	**ICC (Null)**	**ICC (Full)**	**Var (Null)**	**Var (Full)**	**PRV**	**ΔAIC**
BegInt Form 1 (23)	0.429	0.196	160.871	51.644	67.9%	-177.813
BegInt Form 2 (23)	0.410	0.167	190.300	54.369	71.4%	-116.701
BegInt Form 1 (30)	0.435	0.178	171.414	44.755	73.9%	-207.971
BegInt Form 2 (30)	0.466	0.027	194.025	5.750	97.0%	-102.262
IntAdv Form 1 (23)	0.323	0.193	128.075	60.192	53.0%	-953.214
IntAdv Form 2 (23)	0.346	0.154	146.717	48.212	67.1%	-170.983
IntAdv Form 1 (30)	0.291	0.128	88.413	28.567	67.7%	-623.212
IntAdv Form 2 (30)	0.379	0.206	131.879	52.963	59.8%	-115.974
BegIntAdv Form 1 (50)	0.394	0.248	133.273	62.781	52.9%	-98.258
BegIntAdv Form 2 (50)	0.364	0.088	119.175	17.950	84.9%	-83.447

5.2.3 Variance Decomposition and ICC

Across all ten exam forms, the null model teacher ICC ranges between 0.291 and 0.466 which are relatively high ICC values [HH07]. The consistency of ICC values in the null model across forms, regardless of difficulty level, is an important finding and indicates that initial clustering at the teacher level is strong.

However, the inclusion of fixed effects produces large reductions in teacher variance across exam forms. Seven out of ten forms show reductions greater than 60%, which is a substantially larger and more consistent effect than what was observed in the California only analysis. After accounting for covariates, residual teacher ICCs are around the range of 0.15 to 0.20. This consistency is a headline finding of the analysis: variation in LOCUS scores (roughly 15 to 20 percent) remains associated with the teacher adminstering the exam. In regards to the two forms that have extremely low full model ICC values (BegInt Form 2 (30) and BegIntAdv Form 2 (50)), we do not have a strong explanation for why these particular forms

are outliers and would caution against over interpreting a single form's residual ICC; the broader pattern is that residual teacher ICCs cluster between 0.15 and 0.20 after fixed effect adjustment.

5.2.4 Model Fit and Comparison

AIC comparisons favor the full model over the null model across the board with ΔAIC values ranging from -83.4 to -953.2. This corroborates the variance reduction findings: the covariates explain a substantial share of between-teacher variance with AIC signaling a better fit [Aka74].

5.2.5 Discussion

A couple findings stand out from the full data analysis. First, between-teacher clustering is strong and consistent in the LOCUS data especially in the null models before any covariates are accounted for. This is consistent with the pattern that was observed in the California only analysis but with much larger sample sizes. The clustering consistency across difficulty levels, form lengths, and population indicates that the teacher administering a LOCUS exam is meaningfully associated with the scores their students produce. What this means is that simple comparisons across classrooms is inaccurate because a substantial share of score variation reflects between-classroom differences before any student-level factors are even considered.

Second, the inclusion of fixed effects does however explain a large and consistent share of between-teacher variance on every form. This is more consistent than the California only analysis and the larger sample sizes allow for more stable estimation. This suggests that the variability observed in California most likely reflects sampling limitations rather than substantial form level differences.

Lastly, after all the covariates are accounted for, residual teacher level ICCs remain

between 0.15 and 0.20. Thus, there remains a component of teacher-level variation in student LOCUS scores that the observed covariates cannot explain. It is important to not interpret this residual ICC as a causal estimate of teacher effects. The LOCUS sample is not random with respect to teachers since schools and teachers self-select into LOCUS administration, and individual observations might be correlated with unobserved factors such as student prior achievement, school resources, and classroom composition. Overall, the residual teacher ICC represents a rough upper limit on how much teachers might matter while also suggesting that some of the differences may still be explained by factors that have not been accounted for. Subsequent sections refine this interpretation further by examining which covariates carry explanatory power and how state effects behave across forms.

5.3 Marginal Variance Decomposition

Section 5.2 showed that year, population, testing phase, and state jointly explain a large share of the teacher variance but the joint model does not tell us which covariate is doing the most work. The full dataset analysis suggests that state is the most consequential and this section is a direct test of that claim. In order to see which covariate is the most dominant, we use marginal PRVs for this decomposition.

5.3.1 Setup

The three models used for the decomposition are described here:

$$Y_{ij} = \beta_0 + \beta_1(\text{School Year}_j) + u_j + \epsilon_{ij}.$$

$$Y_{ij} = \beta_0 + \beta_2(\text{Population}_j) + u_j + \epsilon_{ij}.$$

$$Y_{ij} = \beta_0 + \beta_3(\text{State}_j) + u_j + \epsilon_{ij}.$$

$$Y_{ij} = \beta_0 + \beta_4(\text{Testing Phase}_j) + u_j + \epsilon_{ij}.$$

Each model is fit per form. Marginal PRV is defined as the percent reduction in teacher variance from the null model to each of the single covariate model. The formula for the marginal PRV is defined below:

$$\text{PRV}_X = \frac{\hat{\sigma}^2_{\text{teacher, null}} - \hat{\sigma}^2_{\text{teacher, +X}}}{\hat{\sigma}^2_{\text{teacher, null}}} \times 100$$

where X is School Year, Population, Testing Phase, or State.

5.3.2 Results

Table 5.3 shows the marginal percent reduction in teacher-level variance for each of the four covariates across all ten exam forms. Each PRV column shows the percent reduction in teacher variance when that single covariate is added to the null model, holding the other covariates out of the model. For example, the '+Pop' column reflects the variance reduction when population alone is added to the null model, not when population is added on top of academic year. This process isolates the contribution each covariate could make on its own and avoids the order dependence that could happen when covariates are added sequentially. The null model teacher variance and ICC are included as a reference baseline and the last column contains the percent reduction when all four covariates are included in the model. Marginal PRVs may take on negative values when a single covariate redistributes variance instead of absorbing teacher-level variance.

Table 5.3: Marginal Teacher-Variance Decomposition by Exam Form

Form	**Teacher Var (Null)**	**Teacher ICC (Null)**	**PRV +Year**	**PRV +Pop**	**PRV +State**	**PRV +Phase**	**PRV (Full)**
BegInt Form 1 (23)	160.9	0.4	9.2%	31.6%	61.6%	-3.7%	67.9%
BegInt Form 2 (23)	190.3	0.4	6.7%	22.2%	63.5%	-3.0%	71.4%
BegInt Form 1 (30)	171.4	0.4	-3.6%	35.5%	34.1%	-11.0%	73.9%
BegInt Form 2 (30)	194.0	0.5	-4.4%	67.2%	88.6%	-27.1%	97.0%
IntAdv Form 1 (23)	128.1	0.3	-9.9%	30.9%	41.2%	6.3%	53.0%
IntAdv Form 2 (23)	146.7	0.3	12.9%	-7.6%	75.0%	4.5%	67.1%
IntAdv Form 1 (30)	88.4	0.3	-6.0%	8.6%	39.1%	24.4%	67.7%
IntAdv Form 2 (30)	131.9	0.4	-2.3%	30.3%	43.7%	5.6%	59.8%
BegIntAdv Form 1 (50)	133.3	0.4	14.7%	18.6%	57.8%	8.2%	52.9%
BegIntAdv Form 2 (50)	119.2	0.4	43.6%	10.3%	55.1%	-22.4%	84.9%

Note. Each PRV column reports the percent reduction in teacher-level variance relative to the null model when the indicated covariate is added individually, holding all other covariates out of the model. The final column reports the percent reduction under the full model including all covariates.

5.3.3 Discussion

Across all exam forms, state is the dominant single covariate, with marginal PRV values averaging around 55%. This suggests that the consistent variance reductions observed in the full dataset analysis may be attributed to state mean differences. Thus, what initially appears to be variation between teachers in LOCUS performance can be accounted for by knowing which state the teacher administered the exam in. For LOCUS reporting, this implies that classroom comparisons that do not adjust for state could conflate teacher-level differences when variation in scores might actually reflect geographic/state patterns.

Population of test takers also contributes moderately with marginal PRV values averaging approximately 24%. However, population's contribution is much more variable than state, ranging from slightly negative values to around 67%. The high variability in PRV values may be due to population distributions across forms. The broader implication is that the value of population as a covariate mostly depends on the composition of test takers on a given form.

Year contributes pretty minimally on its own with marginal PRVs averaging approximately 6%. This suggests that when population and location of test administration is accounted for, academic year does very little in absorbing variation. Given that LOCUS has been validated as a measure of conceptual understanding in statistics, this pattern provides indirect evidence that levels of statistical understanding have remained broadly stable over the 2014-2025 administration period.

Testing phase was included to control for students who might have received more statistical instruction and to account for the Form 2 score advantage that was observed in the exploratory data analysis. On its own, phase explains little of the between-teacher variance and in some cases leads to slight increases in estimated teacher variance. This does not mean phase is uninformative but that it serves its job as a useful control that refines variance estimates once more dominant covariates such as state and population are included. This is

consistent with the original motivation of accounting for timings in instructional exposure without requiring student ID information, and suggests that early versus late testing does not independently drive variation in LOCUS scores.

Overall, the marginal decomposition refines the full-dataset interpretation of the residual ICC of approximately 0.15–0.20 observed in Section 5.2. The decomposition suggests that state accounts for the largest portion of variance, while population also contributes for certain exam forms. State's dominant role raises a question that Section 5.4 attempts to answer: if state accounts for much of the between-teacher variation, does it reflect genuine state-level educational characteristics or is it simply a proxy for which teachers, schools, and programs administer LOCUS in each state?

Importantly, it is worth noting that the marginal decomposition bounds the maximum variance a covariate could explain on its own, not the variance in the presence of other covariates. This approach makes more sense then sequentially adding each covariate to the model and then measuring the drop in variance as this would impose an ordering assumption as to what covariates need to be added and in what order. What the decomposition does establish is that the reduction from the null ICC to the residual ICC in the full model reflects the joint adjustment of state, population, year, and phase, with state appearing to play the most substantial marginal role across most exam forms.

5.4 State Effect Stability Analysis

Building off the marginal decomposition, this section further examines the dominance of state as a fixed-effect. If state effects truly reflect characteristics of state educational systems, then the estimated state coefficients should be reasonably consistent across forms. We examine this assumption directly by extracting the estimated state coefficients from the full models in Section 5.2 and look at their stability across the exam forms with sufficient state coverage.

5.4.1 Setup

Not all combinations of state and forms support stable estimation of state effect coefficients. For example, many states contributed only one or two teachers to the dataset which could yield imprecise estimates if included. Similarly, some forms have very narrow state representation; thus, before interpreting estimated state coefficients, it was important to check whether sufficient representation existed within each state-form combination. Forms that were chosen depended on whether multiple states had more than three teachers and at least 50 students. Six forms satisfied this criterion: BegInt Form 1 (23), BegInt Form 2 (23), BegInt Form 1 (30), IntAdv Form 1 (23), IntAdv Form 1 (30), and BegIntAdv Form 1 (50). Among the forms that were kept, seven states had sufficient combined representation: California, Florida, Virginia, Wisconsin, Ohio, New York, and Arizona. California was used as the reference category because it appeared on all ten forms and had the largest overall sample of teachers and students.

The state coefficients analyzed in this section are extracted from the full models specified in Section 5.2. Each state coefficient represents the estimated difference in mean exam score (in percentage points) between a student in that state and a comparable California student, after adjusting for year, population of test taker, and test phase. This analysis focuses on the consistency and pattern of variability of state effects across forms. Thus, stability conclusions do not change depending on which state is used as the reference category.

5.4.2 Results

Table 5.4 reports the estimated state effects, expressed in percentage points relative to California, from the exam forms that meet the coverage threshold. Form labels in the table follow a scheme. For example, BI23-1 denotes BegInt Form 1 with 23 questions, and IA30-1 denotes IntAdv Form 1 with 30 questions. Asterisks next to each estimate denote statistical significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The final column reports each state's

mean estimate across the forms. Figure 5.1 displays the same estimates as a forest plot, with horizontal bars representing 95% confidence intervals and a vertical dashed line indicating California (the reference category, fixed at zero).

Overall, every state shows a positive mean effect relative to California when averaged across all forms. State coefficients also vary substantially across forms within a state. Florida, Ohio, and Wisconsin have positive estimates on all six forms while Arizona, Virginia, and New York show least one negative estimate. This variation within state makes it difficult to make stable state-level interpretations; a state effect that reflects characteristics of a state's educational system would be expected to remain reasonably constant across forms.

In addition, the precision of the estimates is uneven across forms. As Figure 5.1 shows, confidence intervals are tight and mostly to the right of zero on BegInt Form 1 (23) but wide enough to cross zero on most of the other five forms. Statistical significance is mostly on the two 23-item BegInt forms and on the remaining forms most estimates do not reach significance regardless of their magnitude.

Overall, the results are mixed, state effects are directionally positive relative to California on average, but the magnitudes vary considerably within each state across forms.

5.4.3 Discussion

The state effect stability analysis does not contradict the variance reduction results that were attributed to state in Section 5.3; adding state genuinely improves model fit and that statistical contribution is real. However, what this analysis does clarify is the interpretation of that contribution.

One interpretation is that state is absorbing heterogeneity but not necessarily state characteristics. If state effects were to represent characteristics of state educational systems, the estimated effects should show agreement across forms within state, but the estimated effects show the opposite pattern with much variation in magnitude and, in a few cases,

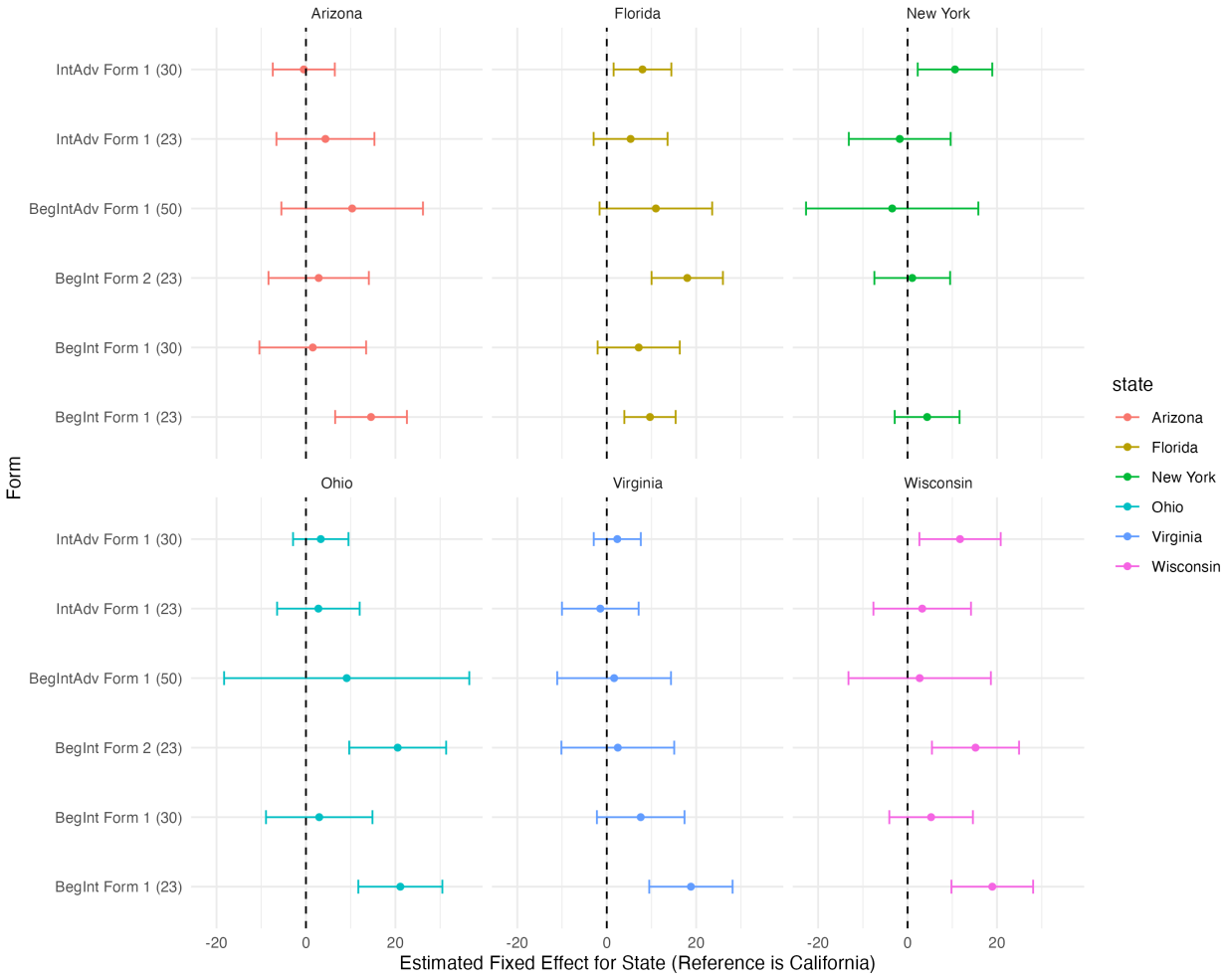

Figure 5.1: Estimated state fixed effects across exam forms. Points represent estimated coefficients relative to the California reference category, with horizontal bars indicating 95% confidence intervals. The dashed vertical line denotes zero effect.

Table 5.4: State-Level Fixed Effects Across Selected Exam Forms

State	BI23-1	BI23-2	BI30-1	IA23-1	IA30-1	BIA50-1	Mean
Arizona	14.55***	2.84	1.52	4.36	-0.48	10.33	5.52
Florida	9.66***	17.99***	7.14	5.33	7.99*	10.98	9.85
New York	4.37	1.04	NA	-1.76	10.60*	-3.44	2.16
Ohio	21.11***	20.51***	2.97	2.78	3.31	9.11	9.97
Virginia	18.82***	2.47	7.59	-1.45	2.35	1.64	5.24
Wisconsin	18.94***	15.19**	5.26	3.27	11.73*	2.70	9.52

Note. Form labels: BI = Beginner/Intermediate; IA = Intermediate/Advanced; BIA = Beginner/Intermediate/Advanced. The number indicates form length (23, 30, or 50 questions), and the final digit indicates form number (e.g., BI23-1 = BegInt Form 1 with 23 questions). Entries report fixed-effect estimates relative to the reference state California. Asterisks indicate statistical significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

sign reversals. The state effects reach high significance on only two forms (BegInt Form 1 and 2, 23 questions) and are rarely significant on other forms. The overall variability in state coefficients suggests that the effects could be driven by factors that are perhaps more associated with which teachers, schools, and programs administered LOCUS in each state for each form. For example, in some states a university program might be driving testing volume while in other states a wider range of schools participate. This claim is consistent with the observed pattern but not formally tested. Direct testing of this claim would require additional information about the criteria by which schools and teachers selected into LOCUS exam participation, which is not present in the LOCUS dataset currently.

However, it is still important to keep in mind the role of statistical power. It is plausible that most states simply do not contribute the necessary number of teachers or samples on a given form to produce stable coefficient estimates despite our attempts to address this issue. Thus, two potential explanations do exist: state effects genuinely reflect local participation patterns or that real state level effects do exist but cannot be estimated precisely due to insufficient teachers per state. With the second explanation, variability in state coefficients

across forms could reflect sampling noise from low power rather than true instability in effects. The current data cannot rule out either explanation and distinguishing between them would simply require a larger number of teachers per state on each form or external validation from another report in which LOCUS state estimates could be compared to.

Regardless of which explanation is ultimately correct, the state effect stability analysis refines the interpretation of the residual teacher ICC. If state is acting as a proxy for local participation patterns, then the residual teacher ICC reflects the between-teacher variation after adjusting for those participation differences. On the other hand, if meaningful state-level effects exist but cannot be estimated precisely, then some of the residual teacher ICC may still contain unmeasured state-level variation. As a result, the teacher level ICC is better interpreted as an adjusted upper bound on between-teacher variation rather than as a causal estimate of teacher instructional effects.

5.5 Test Taker Stratification Analysis

The full dataset model treated population of test takers as a fixed effect. However, a natural extension is to see whether there are teacher level differences depending on the test taking population and for the exam for which they were intended to take. We test that assumption directly: do AP Statistic teachers and Introductory Statistics teachers exhibit different between-teacher variation in their students' LOCUS score?

There are only two forms that were analyzed for this result as they were the only two forms where both populations have sufficient and roughly balanced number of observations: BegInt Form 1 (23) and the IntAdv Form 1 (23). Other forms either had one population almost entirely absent or had heavily imbalanced counts.

For each selected form and population item, both null and full models are fit on the subset of students belonging to that population on that form. The full model retains school year, state, and testing phase as fixed effects but drops population, since the analysis is

already restricted to a single population:

$$Y_{ij} = \beta_0 + \beta_1(\text{School Year}_j) + \beta_2(\text{State}_j) + \beta_3(\text{Testing Phase}_j) + u_j + \epsilon_{ij}.$$

The output of these models for each selected form is shown in Table 5.5.

Table 5.5: Population-Stratified Teacher-Level Variance Decomposition

Form	**Population**	**n Teachers**	**n Students**	**ICC (Null)**	**ICC (Full)**	**PRV**	**ΔAIC**
BegInt Form 1 (23)	AP Statistics	24	1,411	0.269	0.006	98.4%	-49.1
BegInt Form 1 (23)	Intro Statistics	16	1,282	0.387	0.103	82.4%	-33.1
IntAdv Form 1 (23)	AP Statistics	43	6,332	0.212	0.120	55.1%	-741.0
IntAdv Form 1 (23)	Intro Statistics	13	7,226	0.393	0.334	24.5%	-163.3

5.5.1 Discussion

Some patterns that emerge are that Introductory Statistic teachers consistently have higher teacher ICC values than AP Statistic teachers on the same form. More concretely, this means that two students with the same population label, taking the same form in the same school year, state, and testing phase would be expected to score more similarly if their teachers were both AP teachers than if both were Introductory Statistics teachers.

This is an important result as it indicates that institutional and curricular structures potentially influence teacher variation. AP Statistics teachers typically teach within a standardized national curriculum. Introductory Statistics is administered across a wide range of college and university settings with teachers having much more autonomy over how they may operate the classroom. With this understanding, it seems that standardization in curriculum reduces between teacher variation as seen by lower ICC results between AP Statistic teachers and Introductory Statistics students. This result suggests that residual teacher variation in

LOCUS scores could depend on the population context for a given form.

CHAPTER 6

Conclusion

6.1 Key Findings

This section summarizes the main findings of the thesis and the implications it has for the broader LOCUS development team, researchers analyzing LOCUS or similar assessment data, and the broader statistics education community.

First, the ten anchor questions that appear on every LOCUS form correlate moderately to strongly with non-equating performance. Across all ten forms as well, per-question difficulty estimates remain stable when compared across paired Form 1 and Form 2 exams. These properties are essential to a good equating set. For the LOCUS development team, this provides empirical evidence that the equating design holds up under scrutiny. For researchers of LOCUS data, this supports the use of equating set scores as a comparability metric when working across forms.

Second, teacher-level clustering is a strong feature of the LOCUS data. Null model intraclass correlations at the teacher level range from 0.29 to 0.47 across the ten forms, which is well above what should be typically observed for school-level ICCs in academic achievement data [HH07]. This implies that any analysis of LOCUS scores that ignores teacher membership would simply be incorrect. For researchers, this means that hierarchical modeling is not optional when working with LOCUS, it is required. For the LOCUS team, this also means that descriptive reporting at the student level, without acknowledgment of classroom clustering, can be misleading of actual score variation.

Third, LOCUS scores have remained relatively stable over the 2014 to 2025 administration period. Exam year explains little to no teacher-level variance with marginal PRV values averaging around 6%. What this implies indirectly is that the average levels of conceptual statistical understanding, as measured by LOCUS, has not drifted substantially across the administration period. This is an important property for any assessment that is intended to measure a specific ability and the LOCUS exam appears to do so. For the LOCUS team, this is evidence of measurement stability over time. For statistics educators, this stability is what makes cohort comparisons on LOCUS meaningful in the first place.

Fourth, apparent state-level effects are unstable across forms and most likely represent LOCUS participation patterns rather than educational properties of the state itself. The marginal decomposition identified state as a significant contributor of absorbing teacher-level variance. However, the state effect stability analysis showed that estimated state coefficients vary in magnitude across forms within the same state and really only reach statistical significance on two of the six forms analyzed. This pattern is not what one would expect from genuine state-level educational effects. It is worth noting that real state-level effects may be present and the coefficients could be reflecting sampling noise rather than the absence of state effects and we acknowledge that this cannot be ruled out. Thus, the implication for researchers is clear: state fixed effects can reduce teacher variance in statistical models but their coefficients should not be interpreted as estimates of state-level impacts without more evidence. For the LOCUS team, this finding shows that reporting state-level comparisons as claims about the state education systems should be proceeded with caution.

Lastly, after accounting for observable covariates, roughly 15 to 20 percent of score variation remains associated with the teacher administering the exam and this variation very much depends on curricular context. Across all ten exams, residual teacher ICCs fall in the range of approximately 0.15 to 0.20 once school year, population, state, and testing phase are included in the final model. This means that two students with the same population label, taking the same form in the same school year, state, and testing time frame, can still

be expected to have different scores on the LOCUS exam depending on which teacher administered the exam. The population-stratified analysis qualifies this further; AP Statistics teachers show lower teacher ICCs than Introductory Statistics teachers. This can potentially be attributed to the fact that standardization in curriculum, which AP Statistics teachers must teach under, might be associated with lower between-teacher variation. Thus, the residual teacher ICC finding could be a number that also depends on instructional setting. For statistics educators, the AP versus Introductory Statistics finding is a result worth taking seriously as it offers potential empirical support for the idea that standardized curriculum may reduce between-teacher variation. The LOCUS data offers possibilities for examining that question further.

6.2 Limitations and Future Work

The most important limitation is the absence of a student level prior achievement measure. The natural extension is therefore to incorporate measures of student prior achievement by potentially linking to external academic records. If a prior achievement covariate were available and the residual teacher ICC was substantially reduced after including it in the model, that would suggest that student selection was a main driver of the results rather than the teacher's actual teaching abilities. If the residual ICC did not change after including the covariate, that would strengthen the case for a meaningful teacher effect. Another related extension would be obtaining unique student IDs so that Form 1 and Form 2 scores could be paired between the same students; this pairing would allow for analysis of score changes within students. Relatedly, because teacher-level effects are important, future studies of students' understanding of statistics should consider teacher-level variables such as experience teaching statistics, type of school (public vs. private), or class size. The absence of this sort of information in the LOCUS dataset limits the ability to explain why teachers differ but collecting and including teacher-level data could help remedy this limitation.

Another consideration concerns school and district level variation. The empirical teacher-to-school ratio in the LOCUS data is close to 1:1 in most exam forms, meaning that most schools only had observations from a single teacher. Thus, the teacher random intercept is essentially capturing school level variation as well since we expect most middle school and high schools to only have one teacher instructing statistics. In this sense, the absence of school level clustering is more of a structural feature of the data rather than a limitation. On the other hand, a more meaningful limitation is the absence of district level information. Without district identifiers, it is not possible to assess whether district level policies or resources are contributing to observed variation. As the LOCUS dataset grows, ensuring district level representation would make it possible to investigate and model their effects explicitly.

Another closely related limitation is that state was modeled as a fixed effect rather than as a random level. Expanding the LOCUS dataset to include more states and ensuring that each state contributes multiple teachers, would allow state to be modeled as a third level with students within teachers within states. This would also make it possible to investigate whether the residual teacher ICC of approximately 0.15 to 0.20 is consistent across states or whether it varies with state-level factors.

It is important to note as well that the LOCUS dataset is not nationally representative. Selection into LOCUS administration is non-random, with some states and student populations heavily overrepresented, and generalization to assessments other than LOCUS would require additional assumptions that this analysis does not test.

Several additional modeling choices simplify the analysis in ways that are worth acknowledging. The testing phase variable captures when in the academic year an exam was given and is not meant to act as a proxy for pre/post test examination. The phase covariate is supposed to act more as a control for administration timing during the academic cycle as students who took the test later on in the school year may have received academic instruction which could influence scores. Additionally, all models in the full dataset analysis assume

that student level residuals are independent conditional on the teacher random intercept. In practice, students within the same classroom may remain correlated and a more flexible specification would relax this assumption at the cost of additional parameters. However, the variance components estimated in the current analysis show that student-level variance is consistently larger than teacher-level variance so the conclusions regarding between-teacher variance are not likely to be affected by the conditional independence assumption.

6.3 Conclusion

In conclusion, this thesis offers evidence that the teacher administering a LOCUS exam is meaningfully associated with the scores that students produce on it. A naive analysis of LOCUS scores that ignored teacher membership would understate its importance in this educational setting. Overall, the findings of this thesis provide a more complete understanding of the factors associated with students' conceptual understanding of statistics and highlight the importance of careful analysis when interpreting assessment outcomes in an educational context.

APPENDIX A

Table A.1: Distribution of teachers, schools, and students by state for BegInt Form 1 (23).

State	**Teachers**	**Schools**	**Students**
California	27	14	2,058
Florida	11	10	642
New York	6	5	898
Texas	5	5	319
Arizona	4	4	5,909
Wisconsin	3	3	426
Virginia	3	3	368
Ohio	3	3	192
New Jersey	2	2	63
Washington	2	2	43
Alabama	2	2	25
Maryland	2	2	23
Illinois	1	1	790
South Carolina	1	1	136
Nevada	1	1	59
Colorado	1	1	39
Missouri	1	1	34
Michigan	1	1	32
North Carolina	1	1	28
Georgia	1	1	24
Pennsylvania	1	1	23
Tennessee	1	1	12
Massachusetts	1	1	11

Table A.2: Distribution of teachers, schools, and students by state for BegInt Form 2 (23).

State	**Teachers**	**Schools**	**Students**
California	27	14	1,828
Florida	6	5	326
New York	5	4	727
Texas	4	4	305
Wisconsin	3	3	290
Ohio	3	3	149
Washington	3	3	54
Arizona	2	2	5,303
Virginia	2	2	261
New Jersey	2	2	241
Colorado	2	2	97
Nevada	2	2	53
Missouri	1	1	437
Maryland	1	1	32
Alabama	1	1	28
Massachusetts	1	1	17
Illinois	1	1	13
Kentucky	1	1	12
Idaho	1	1	7
Indiana	1	1	6

Table A.3: Distribution of teachers, schools, and students by state for BegInt Form 1 (30).

State	**Teachers**	**Schools**	**Students**
California	7	5	1,275
Wisconsin	4	4	339
Florida	4	4	330
Virginia	3	2	409
Massachusetts	3	3	60
Ohio	2	2	86
Georgia	2	2	47
Alabama	2	2	42
Arizona	2	2	42
Missouri	2	2	28
South Carolina	1	1	232
Texas	1	1	129
Illinois	1	1	107
Maryland	1	1	48
Michigan	1	1	30
New Jersey	1	1	28
Louisiana	1	1	19
Pennsylvania	1	1	16
Washington	1	1	14
Indiana	1	1	9

Table A.4: Distribution of teachers, schools, and students by state for BegInt Form 2 (30).

State	**Teachers**	**Schools**	**Students**
California	5	3	885
Florida	2	2	217
Alabama	2	2	49
South Carolina	1	1	161
Texas	1	1	62
Michigan	1	1	50
Massachusetts	1	1	18
Washington	1	1	7

Table A.5: Distribution of teachers, schools, and students by state for IntAdv Form 1 (23).

State	**Teachers**	**Schools**	**Students**
Florida	8	8	1,477
California	7	7	7,037
Virginia	7	6	722
Massachusetts	6	6	192
Ohio	5	5	414
Arizona	3	3	237
Wisconsin	3	3	184
New York	3	3	88
New Jersey	2	2	292
Texas	2	2	261
Michigan	2	2	176
Illinois	1	1	3,257
Missouri	1	1	559
Indiana	1	1	180
Minnesota	1	1	167
Pennsylvania	1	1	52
Iowa	1	1	48
South Carolina	1	1	38
Colorado	1	1	27
Montana	1	1	26
Alabama	1	1	22
Maryland	1	1	17
Idaho	1	1	17

Table A.6: Distribution of teachers, schools, and students by state for IntAdv Form 2 (23).

State	**Teachers**	**Schools**	**Students**
California	5	5	2,689
Virginia	4	4	209
Georgia	3	3	97
Florida	2	2	189
Arizona	2	2	164
Ohio	2	2	130
Wisconsin	2	2	124
New Jersey	1	1	268
Michigan	1	1	257
Texas	1	1	113
South Carolina	1	1	70
Montana	1	1	26
North Carolina	1	1	23
Missouri	1	1	20
New York	1	1	19
Louisiana	1	1	14
Alabama	1	1	12
Maryland	1	1	12

Table A.7: Distribution of teachers, schools, and students by state for IntAdv Form 1 (30).

State	**Teachers**	**Schools**	**Students**
California	10	10	1,878
Virginia	9	8	911
Florida	5	4	600
Ohio	5	5	509
Arizona	4	4	187
New York	3	3	694
Michigan	3	3	387
Illinois	3	3	74
South Carolina	3	3	61
Georgia	2	2	101
Indiana	2	2	101
Massachusetts	2	2	87
Alabama	2	2	77
Missouri	2	2	77
Wisconsin	2	2	74
Texas	2	2	56
Pennsylvania	2	2	46
Nevada	2	2	36
Maryland	2	2	29
New Jersey	1	1	179
Minnesota	1	1	81
Delaware	1	1	34
Connecticut	1	1	32
Tennessee	1	1	25

State	Teachers	Schools	Students
Mississippi	1	1	22
North Carolina	1	1	19
Washington	1	1	10
Colorado	1	1	8

Table A.8: Distribution of teachers, schools, and students by state for IntAdv Form 2 (30).

State	**Teachers**	**Schools**	**Students**
Virginia	5	4	301
California	4	4	862
Florida	4	4	373
Ohio	3	3	127
Michigan	2	2	175
Missouri	2	2	68
Arizona	2	2	66
Wisconsin	2	2	43
Illinois	2	2	23
New York	1	1	485
Minnesota	1	1	94
Massachusetts	1	1	52
North Carolina	1	1	44
Colorado	1	1	29
Alabama	1	1	26
New Jersey	1	1	19
Texas	1	1	18
Tennessee	1	1	14
Nevada	1	1	11

Table A.9: Distribution of teachers, schools, and students by state for BegIntAdv Form 1 (50).

State	**Teachers**	**Schools**	**Students**
Florida	4	4	406
California	4	4	279
Georgia	4	4	218
Virginia	3	3	240
Wisconsin	3	2	219
Minnesota	2	2	161
Arizona	2	2	47
North Carolina	1	1	73
Ohio	1	1	65
New Jersey	1	1	59
New York	1	1	44
Pennsylvania	1	1	30
Connecticut	1	1	16
Massachusetts	1	1	13
Alabama	1	1	9
Washington	1	1	9

Table A.10: Distribution of teachers, schools, and students by state for BegIntAdv Form 2 (50).

State	**Teachers**	**Schools**	**Students**
California	3	3	146
Florida	2	2	383
Georgia	2	2	128
Wisconsin	1	1	78
New York	1	1	41
Virginia	1	1	39
Indiana	1	1	39
Pennsylvania	1	1	17

REFERENCES

[Aka74] Hirotugu Akaike. "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control*, **19**(6):716–723, 1974.

[Ang84] William H. Angoff. *Scales, Norms, and Equivalent Scores*. Educational Testing Service, Princeton, NJ, 1984.

[BMB15] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software*, **67**(1):1–48, 2015.

[GH07] Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2007.

[HH07] Larry V. Hedges and Edward C. Hedberg. "Intraclass correlation values for planning group-randomized trials in education." *Educational Evaluation and Policy Analysis*, **29**(1):60–87, 2007.

[JdH23] Tim Jacobbe, Bob delMas, Brad Hartlaub, Jeff Haberstroh, Catherine Case, Steven Foti, and Douglas Whitaker. "Establishing the Validity and Reliability of the LOCUS Assessments." *Numeracy*, **16**(1):Article 5, 2023.

[LH21] Eunsoo Lee and Sehee Hong. "Adequate Sample Sizes for a Three-Level Growth Model." *Frontiers in Psychology*, **12**:685496, 2021.

[Liv04] Samuel A. Livingston. "Equating Test Scores (Without IRT)." Technical report, Educational Testing Service, Princeton, NJ, 2004.

[RB02] Stephen W. Raudenbush and Anthony S. Bryk. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications, 2 edition, 2002.

[WFJ15] Douglas Whitaker, Steven Foti, and Tim Jacobbe. "The Levels of Conceptual Understanding in Statistics (LOCUS) Project: Results of the Pilot Study." *Numeracy*, **8**(2):Article 3, 2015.