Boston College
Lynch School of Education

Department of
Measurement, Evaluation, Statistics, and Assessment

A COMPARISON OF METHODS FOR ESTIMATING STATE SUBGROUP PERFORMANCE ON THE NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

Dissertation
by

DAVID BAMAT

submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

March 2021

© Copyright by David Bamat
2021

Abstract

A Comparison of Methods for Estimating State Subgroup Performance on the National Assessment of Educational Progress

David Bamat

Dr. Henry Braun, Dissertation Chair

The State NAEP program only reports the mean achievement estimate of a subgroup within a given state if it samples at least 62 students who identify with the subgroup. Since some subgroups of students constitute small proportions of certain states' general student populations, these low-incidence groups of students are seldom sufficiently sampled to meet this rule-of-62 requirement. As a result, education researchers and policymakers are frequently left without a full understanding of how states are supporting the learning and achievement of different subgroups of students.

Using grade 8 mathematics results in 2015, this dissertation addresses the problem by comparing the performance of three different techniques in predicting mean subgroup achievement on NAEP. The methodology involves simulating scenarios in which subgroup samples greater or equal to 62 are treated as not available for calculating mean achievement estimates. These techniques comprise an adaptation of Multivariate Imputation by Chained Equations (MICE), a common form of Small Area Estimation known as the Fay-Herriot model (FH), and a Cross-Survey analysis approach that emphasizes flexibility in model specification, referred to as Flexible Cross-Survey Analysis (FLEX CS) in this study. Data used for the prediction study include public-use state-level estimates of mean subgroup achievement on NAEP, restricted-use student-level achievement data on NAEP, public-use state-level administrative data from Education Week, the Common Core of Data, the U.S. Census Bureau,

and public-use district-level achievement data in NAEP-referenced units from the Stanford Education Data Archive.

To evaluate the accuracy of the techniques, a weighted measure of Mean Absolute Error and a coverage indicator quantify differences between predicted and target values. To evaluate whether a technique could be recommended for use in practice, accuracy measures for each technique are compared to benchmark values established as markers of successful prediction based on results from a simulation analysis with example NAEP data.

Results indicate that both the FH and FLEX CS techniques may be suitable for use in practice and that the FH technique is particularly appealing. However, before definitive recommendations are made, the analyses from this dissertation should be conducted employing math achievement data from other years, as well as data from NAEP Reading.

Acknowledgements

I would like to thank my advisor and dissertation chair, Dr. Henry Braun, for his wisdom, guidance and patience throughout this process.

I would also like to thank the other members of my dissertation committee, Dr. Laura O'Dwyer and Dr. Matthias von Davier, both of whom gave me important feedback and greatly improved the quality of this dissertation.

In addition, I would like the MESA department and Boston College for the opportunity to pursue this PhD and grow as an Education researcher.

Finally, I would like to thank my wife Liz for her unwavering support and encouragement to continue working on this dissertation until the very end.

Contents

List of Tables

List of Figures

Chapter 1: Introduction

Overview

State NAEP (National Assessment of Educational Progress) is an American assessment program that administers biennial achievement tests in reading and mathematics, in grades 4 and 8, to representative samples of students from each of the 50 states, plus the District of Columbia and students in schools managed by the Department of Defense. The NAEP program has two major goals: to compare student achievement among states and other jurisdictions, and to track changes in achievement of fourth-, eight-, and twelfth-graders over time (U.S. Department of Education, 2015a).[1] One of State NAEP's greater affordances is its disaggregation of student test results by demographic subgroup, currently comprising eighteen different subgroups. This disaggregation allows policymakers and researchers to gain a sense of how states are supporting the learning of different subgroups, including underserved and underperforming groups of students.

Unfortunately, State NAEP does not report on the achievement of all eighteen subgroups for each of the fifty states. As a policy, the program only reports subgroup results if it samples at least 62 students who identify with the subgroup within any given state (Elliott & Phillips, 2004; Chromy, Finker & Horvitz, 2004). Since some subgroups of students represent small proportions of certain states' general student populations (e.g., Black students in Vermont), these low-incidence groups of students within certain states are insufficiently sampled to meet the requirement. As a result, while the NAEP program publishes estimates of mean achievement and standard errors for subgroups of students that are more common within states (e.g., White students in Vermont, Hispanic students in California), blank spaces or symbols demarcating

[1] Since 2003, State NAEP tests have been administered to fourth- and eighth-graders every two years (biennially), but to twelfth-graders every four years (U.S. Department of Education, 2017).

omitted results replace would-be achievement estimates for many minority populations within states in NAEP publications.

This study addresses this limitation by comparing the performance of different techniques in predicting the mean math achievement of subgroups on State NAEP. Ultimately, it aims to answer whether it may be justifiable to apply one or more of the techniques under examination to the estimation of mean subgroup achievement on State NAEP when direct estimation is impermissible because of insufficient sample size.

Background

This research is motivated by a demand for estimates of subgroup achievement. Meeting such demand advances the twin goals of NAEP, to compare student achievement in states and other jurisdictions, and to track changes in achievement. Education officials and policymakers want to know how students from different backgrounds are performing and progressing academically (Musu-Gillette et al., 2016). Because NAEP is unable to produce estimates for all subgroups within each state, officials and policymakers receive an incomplete picture of the standing and progress of all groups of students across states.

Another motivation for this research is the prospect of learning important information about the relative strengths and limitations of the techniques and data used for predicting mean subgroup achievement, including which kinds of extant administrative data are most helpful for predicting mean subgroup achievement.[2] While the techniques applied in this dissertation are widely used across different fields of study, they have not been applied to the prediction of mean subgroup achievement on State NAEP.

[2] Administrative data refer to data that are not originally collected for the purpose of estimating parameters of interest. In this study, these are data that *are not* collected by organizations charged with implementing the NAEP program, such as the U.S. Census Bureau, but may nonetheless be helpful for estimating NAEP achievement.

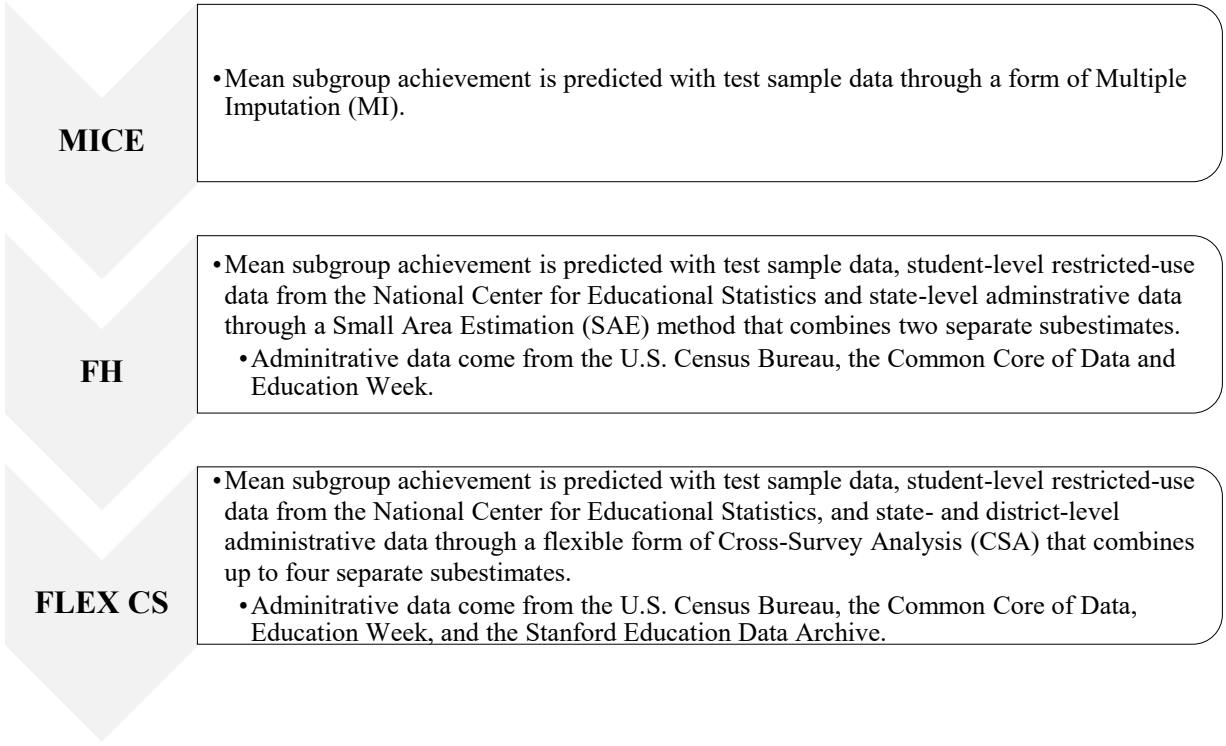
Overview of the Methods Used in this Study

Three separate and progressively more complex analytic techniques are used to predict the mean math achievement of subgroups across states. The predicted values from each technique are then compared with NAEP-reported estimates, which play the role of target values, to gauge each technique's predictive accuracy. Hence, comparisons are applied to cells from the test sample dataset (i.e., grade 8 mean math achievement across states and subgroups in 2015) where NAEP estimates are available. The measures of accuracy are summarized for each technique both *across* subgroups of interest, as well as *by* subgroup, which permits conclusions with respect to whether the relative predictive accuracy of the techniques vary systematically by subgroup. The subgroups of interest, for which predicted estimates of mean math achievement are compared to NAEP-reported estimates, are the subgroups from the test sample with incomplete reporting across states.

The first technique is an adaptation of a Multiple Imputation (MI) method known as Multivariate Imputation by Chained Equations (MICE). The second is a Small Area Estimation (SAE) technique known as the Fay-Herriot (FH) model. The third technique is a more novel approach that emphasizes flexibility in model specification and combines features of the MI and SAE approaches, an approach referred to in this dissertation as Flexible Cross-Survey Analysis (FLEX CS). As depicted in Figure 1.1, the techniques become progressively more complex in terms of the sources of data that they require for prediction as well as the manner in which estimates (subestimates) are combined for predicting the mean math achievement of subgroups.[3]

[3] It should be noted that the meaning of *progressive complexity* here to describe a succession of techniques differs from how it is often used in the research literature to describe regression models with successively larger sets of predictor variables. By contrast, the progressive complexity across techniques in this study is related more to additional *sources* of data that the techniques require, as well as the manner by which estimates are formed from subestimates across techniques.

***Figure 1.1*:** Progressive complexity of prediction techniques



Measuring Performance (Predictive Accuracy)

To evaluate the accuracy of the three techniques in predicting subgroup achievement on State NAEP, estimates of mean math achievement produced from the techniques are compared to target values representing NAEP-reported estimates of mean math achievement. Measures of accuracy are based on two statistics. The first, weighted Mean Absolute Error (wMAE), is a weighted measure of the distance between estimates produced from the techniques under study and NAEP-reported estimates. The second, coverage, is the frequency with which estimates produced from the techniques under study are located within target intervals associated with NAEP-reported estimates.

Techniques for Predicting Mean Math Achievement of Subgroups

Multiple Imputation (MI)

The first technique used for predicting mean subgroup achievement, an adaptation of Multivariate Imputation by Chained Equations (MICE), is a form of Multiple Imputation (MI).[4] MI is an example of a *modern* missing data analysis approach. Such approaches are preferred to traditional ones such as mean substitution as they more effectively preserve the relationships between variables from the original dataset (Dempster, Laird & Rubin, 1977; Schafer & Graham, 2002). Unlike single-imputation techniques, MI techniques account for the uncertainty in imputations by creating multiple predictions for each missing value through a sequence of random draws from conditional distributions (Azur et al., 2011). As a result, while the imputed values represent plausible values for those that are missing, the imputed values are designed to differ across the imputed data sets.

Multivariate Imputation by Chained Equations (MICE)

The MICE technique represents a fully conditional specification (FCS) approach to imputation whereby the imputation model is specified on a variable-by-variable basis by a set of conditional densities, one for each incomplete variable. In this approach, variables with missing data are successively regressed on select predictor variables from the dataset, and the imputed values represent a series of random draws from the conditional distributions estimated by the predictor variables (van Buuren & Groothuis-Oudshoorn, 2011). The data used for prediction with MICE in this study represent state-level estimates of mean math achievement for subgroups across states.

[4] This technique is referred to as an *adaptation* because the mice technique is typically applied to actual missing data situations. By contrast, the technique is used to predict observed values in this study. More details on this procedure are provided in Chapter 3.

Small Area Estimation (SAE)

Small Area Estimation is a technique for estimating the parameters of subpopulations when the samples available for producing direct estimates (i.e., design-based estimates) of the subpopulation parameters are insufficiently large, which in turn does not permit the calculation of estimates with a desired level of precision. The term "small area" typically refers to smaller geographic areas or minority populations that form part of larger sampled areas and populations. While sampling designs usually result in sample sizes large enough for reasonably precise inference for larger areas and populations, they do not always result in sufficient sample sizes for all subpopulations of interest and, hence, the impetus for the SAE framework (Rao & Molina, 2015; Pfefferman, 2002).

Estimation within the SAE framework is characterized by borrowing information from administrative data related to subpopulation parameters of interest to supplement direct estimates. SAE techniques principally differ from missing data techniques, including multiple imputation, as SAE approaches are not typically used to estimate data that are missing. Instead, SAE techniques are usually implemented to improve direct estimates that are calculated imprecisely (Rao & Molina, 2015; Pfefferman, 2002). Another important difference between imputation and SAE techniques is that the latter require use of administrative data *supplemental* to the original data sample collected for estimation (Rao & Molina, 2015), whereas imputation techniques use *in-sample* data to estimate missing values (Graham, 2009**)**.

The Fay-Herriot Model (FH)

In this study, an area-level model known as the Fay-Herriot model (FH) is used for prediction of mean subgroup achievement (Molina & Marhuenda, 2015; Pfefferman, 2002). The term "area-level" refers to the fact that the administrative data used to supplement direct

estimates are from the same unit level of inference. In the context of this study, state-level variables are used to calculate model-based estimates of mean subgroup achievement across states, which supplement direct estimates computed from student-level data. In the FH technique, final estimates are referred to as *Empirical Best Linear Unbiased Predictors* (EBLUPs). The EBLUP calculated with the FH approach is a precision-weighted combination of a direct estimate of a parameter and a regression estimator (i.e., model-based estimate) of the parameter. The EBLUP is weighted toward the estimate calculated with greater estimated precision. For instance, the greater the uncertainty of the regression estimate relative to the uncertainty of the direct estimate, the more the EBLUP shrinks toward the direct estimate.

Cross-Survey Analysis (CSA)

Cross-Survey Analysis (CSA) refers to the combined analysis of data from different surveys (Magadin de Kramer, 2016). The principal distinction between CSA and SAE is the latter's emphasis on updating a direct estimate; improving the efficiency of an estimate that is design-based. The emphasis of CSA, on the other hand, is on combining data from different sources (surveys) as a means of improving the accuracy of prediction. CSA can be conceived as an analog to Meta-Analysis, with an emphasis on combining observational data collected from different surveys.

Flexible Cross-Survey Analysis (FLEX CS)

The last technique used for estimating mean subgroup achievement, FLEX CS, is a more novel approach that combines features of the MICE and SAE techniques. FLEX CS permits the researcher to select only the variables from the original data file that are presumably most helpful for predicting values in outcome variables, while at the same time allowing the researcher to use administrative data external to the original data file that are useful for prediction. This approach

expands on the flexibility offered by MICE, in which select variables from the original data file are used to impute values on a variable-by-variable basis, while simultaneously borrowing useful predictive data from other surveys.

The technique is described as a form of CSA in this dissertation since the estimates of mean math achievement that are produced from this technique are the combinations of other estimates, or *subestimates,* which are computed from different survey data, judiciously selected by the researcher. The subestimates that form the FLEX CS estimates are computed from *up to* four different techniques. Emphasis is placed on "up to," as the final estimates of mean subgroup achievement within states are not required to be formed from the same subestimates. Instead, they are formed from subestimates that are more defensibly presumed to be accurate estimators of the mean math achievement of a particular state's subgroup, given characteristics of the data from which the subestimates are formed. Final FLEX CS estimates, in this study, are precision-weighted averages of the subestimates that meet certain criteria for contributing to FLEX CS estimates. In addition to state- and student-level data, district-level data from the Stanford Education Data Archive (Reardon et al., 2017) are used for computing subestimates of mean math achievement in the FLEX CS approach.

Evaluating the Techniques

The coverage statistic calculated in this study plays the important role of guiding the decision about whether a technique in this study could be recommended for use in practice—for instance, by researchers carrying out secondary analyses of NAEP data. Researchers should have confidence in a technique for which the vast majority of differences between predicted and target values of mean math achievement, both across and by subgroup of interest, are negligible in size.

In this study, a simulation analysis, described in greater detail in chapter 3, is conducted to establish coverage rate criteria to represent markers of successful prediction.

The other measure, weighted Mean Absolute Error (wMAE), is better suited to help determine which techniques perform better and worse among the three techniques. In the event one or more techniques meet criteria to be considered suitable for use in practice, wMAE statistics help determine which of the techniques performs best in terms of its ability to accurately predict mean subgroup achievement compared to the other techniques.

About Subgroup Reporting on State NAEP

The passage of the No Child Left Behind Act (NCLB) in 2002 ushered in important changes to the NAEP program. In addition to requiring that all fifty states participate in NAEP testing in both reading and math at grades 4 and 8 (Bourque, 2004), NCLB required that reporting of test results be disaggregated by demographic subgroup. Currently, State NAEP disaggregates and reports results based on six separate demographic categories—including race and ethnicity, gender, eligibility for the federal free- and reduced-price school lunch program, highest level of parental education, learning disability status, and English learner status. Each category includes between two to six subgroups that together equal eighteen separate subgroups for which the NAEP program reports estimates of mean achievement (U.S. Department of Education, 2018).

In many instances however, the NAEP program does not report results of all subgroups for each state. As previously stated, the program only reports subgroup results if it samples at least 62 students who identify with the subgroup. This policy was decided by the National Center for Education Statistics (NCES) after NCES researchers determined that 62 was the minimum sample size they would need to detect an effect size of 0.50, with power of 0.50, a 0.05 level of

significance, and a design effect of approximately 2.00 when drawing comparisons between different groups of students (Elliott & Phillips, 2004).

In practice, this means that while the NAEP program has been able to sample enough students to report on subgroups nationally (e.g., estimate of mean math achievement of Black students across the United States), the program is frequently unable to report on mean achievement of subgroups for jurisdictions where subgroup members are uncommon (e.g., Black students in Vermont). To expand on this example, when NAEP randomly samples schools and students in Vermont to take a test, they seldom test at least 62 students who identify as Black because of the low-incidence of Black students in the state of Vermont.

Research Purpose and Research Questions

Research Purpose

The purpose of this dissertation is to ascertain whether one or more of three techniques are suitable for estimating mean subgroup achievement on State NAEP. Ultimately, it aims to answer whether one or more of the techniques can justifiably be applied to estimates of mean subgroup achievement on State NAEP when direct estimation is impermissible because of insufficient sampling.

Research Questions[5]

This dissertation aims to answer the following questions:

[5] It should be noted that research findings from this study cannot be used to make claims about the efficacy of applying the techniques to NAEP achievement data *in general*, as analysis is conducted on one of several possible test samples (in this study, grade 8 math in 2015). The findings can nonetheless serve as a set of initial evidence that informs follow-up and expanded research on this topic, which may ultimately lead to more substantiated claims about the generalizability of the techniques under evaluation.

- Is it reasonable, based on benchmarks established through a simulation analysis, to use any of the techniques examined in this study to estimate subgroup math achievement on State NAEP when sample sizes do not permit direct estimation?
- How do the techniques compare with respect to maximizing accuracy, according to accuracy measures used in this study (weighted Mean Absolute Error and coverage)?
- How do the techniques vary in their ability to predict achievement per subgroup?

Research Design and Methods

The comparison of estimation techniques in this study begins with an evaluation of the performance of Multivariate Imputation by Chained Equations (MICE), followed by the Fay-Herriot model (FH), and then Flexible Cross-Survey Analysis (FLEX CS). Each technique is successively more complex in terms of the data entered into the models and the manner in which estimates are constructed, and it is presumed that the added complexity improves predictive accuracy. Evaluation of the techniques is based on predictive accuracy—in general, how close the predicted values produced from each of the three techniques come to values that are reported by the NAEP program. The set of data on which these analyses are conducted (i.e., the test sample) come from the National Center for Education Statistics and represent mean subgroup achievement by state on the grade 8 mathematics assessment in 2015. The dimensions of this dataset are 50 rows (i.e., observations) by 18 columns. The former corresponds to the 50 American states and the latter corresponds to the 18 subgroups for which State NAEP disaggregates and attempts to report results. In this test sample, 124 of 900 mean achievement values are not reported. The subgroups include:

- Two categories related to the national free or reduced lunch program:
    - Eligible (E), Ineligible (I)

- Four categories related to parental level of education:
    - Did not finish high school (NHS), Graduated high school (HS), Some education after high school (SBA), Graduated college (BA)
- Six categories related to race and ethnicity:
    - White (W), Black (B), Hispanic (H), Asian/Pacific Islander (API), American Indian/Alaskan Native (AIAN), Two or More Races (TP)
- Two categories related to English language proficiency:
    - English language learner (EL), Not English language learner (NEL)
- Two categories related to learning disabilities:
    - Student with disability (SWD), Not student with disability (NSWD)
- Gender:
    - Male (M), Female (F)

Different sources of data and variables are used to support prediction with the three separate techniques.[6] In the MICE procedure, variables from the test sample, with values representing mean subgroup achievement by state on the grade 8 Mathematics assessment in 2015, are incorporated into a series of regression analyses as predictor variables. In the FH procedure, student-level achievement data from a restricted-use NCES database are used to calculate direct estimates, which are combined with synthetic-regression estimates predicted from state-level administrative data representing demographic and school-quality factors from the Common Core of Data (U.S. Department of Education, 2020a), American Community

[6] In a sense, the performance of the prediction techniques under evaluation are not *directly* compared to one another, as they do not draw on the same exact sources of data or variables for prediction. The different data thus represent a potential confounding factor. Put differently, evaluation of the techniques does not include a deliberate effort to parse the utility of the data from the utility of the techniques. The techniques under evaluation, including the data they incorporate, are nonetheless used for the same objective and judgement concerning the predictive performance of these techniques is based on common criteria.

Survey (ACS; U.S. Census Bureau, 2018), and Education Week (Education Week Research Center, 2015). In the FLEX CS procedure, in addition to the administrative data used in the FH approach, district-level achievement data in NAEP-referenced units from the Stanford Educational Data Archive (Reardon et al., 2017) are used for prediction.

Significance of the Study

Results from NAEP can be incredibly useful to researchers and policymakers. The NAEP assessment, which was first administered nationally in 1969 and then to students from all fifty states beginning in 2003, offers the only common metric of achievement on which representative samples of students from all fifty states can be compared (Lapointe, 2004; Olkin, 2004). Results from State NAEP also afford researchers and policymakers the opportunity to gain a sense of how states are supporting the learning of different demographic subgroups, including historically underserved and underperforming groups of students such as low-income, Black, Hispanic, and American Indian students. As such, State NAEP results permit inference, albeit not causal, concerning which states and sets of policies might best support student learning—both for students, in general, as well as for particular demographic groups of students.

Still, the picture painted by State NAEP regarding the learning and achievement of students across states is incomplete, particularly because the NAEP program seldom reports on the mean achievement of low incidence subgroups of students in certain states. This dissertation attempts to answer whether a more complete, yet accurate, picture can be provided through the application of one or more prediction techniques. Specifically, the dissertation seeks to answer whether it might be advisable to use one or more of three separate techniques to estimate mean subgroup achievement on State NAEP when direct estimation is impermissible because of insufficient sample sizes.

If the application of one or more of the techniques under study is justifiable, then a clearer picture of state and subgroup achievement can be provided. Although clearer State NAEP results still do not support causal inference, they offer the opportunity for more accurate inferences. Full reporting of subgroup achievement on State NAEP gives researchers and policymakers improved indications of which states and sets of policies best support the learning and achievement of students, including the learning and achievement of underserved and underperforming groups of students. Most importantly, improved indications foster opportunity for researchers and policymakers to draw important lessons from promising states and sets of policies.

The idea that researchers and policymakers might use indirect estimates (e.g., model-based estimates) of mean subgroup achievement computed from one or more of the techniques under study is conceivable. NCES has published "Full Population Estimates" (FPEs) since 2005, which are model-based estimates of mean achievement adjusted for variation in the degree to which jurisdictions have excluded students with learning disabilities and English learners from NAEP testing (U.S. Department of Education, 2020b). In addition, several federal agencies already use and report model-based parameter estimates when direct estimation is infeasible (Czajka, 2016).

Another tangential benefit of this research is the opportunity to glean important information about the relative efficacy of the techniques under examination in their ability to predict mean subgroup achievement. This research, for instance, can provide insight into which types of administrative data are most helpful for predicting the mean achievement of different subgroups. Revealing information about the relationship between achievement and particular variables invites the opportunity for informed follow-up inquiry regarding factors associated

with learning and achievement, including the learning and achievement of particular subgroups of interest.

Remainder of Dissertation

Chapter 2 details the rationale for this study, including rationale for the proposed research methods and data used for prediction. Chapter 3 describes the research design and methods in greater depth. Chapter 4 describes analysis results. Chapter 5 includes a review of the findings and a discussion of the dissertation's limitations. Finally, the appendices includes a general description of the technical steps undertaken in R and Stata to calculate estimates of mean math achievement, tables demonstrating mean achievement estimates by state and technique, and a series of plots that support the interpretation of results. The statistical code used for each analysis described in the dissertation is provided on the author's GitHub page.

Chapter 2: Review of the Literature

The Case for Studying NAEP

While researchers and policymakers garner a general understanding of the achievement of different groups of students through state-administered tests or college-entrance exams (e.g., the SATs), none of the results from these tests offer the same information as results from NAEP. The NAEP assessment offers the only common metric of achievement on which representative samples of students from all fifty states can be compared (Lapointe, 2004; Olkin, 2004). Unlike college entrance exams, NAEP provides achievement results from samples of students from each state that are demographically characteristic of the state. On the other hand, tests like the SATs or ACTs provide achievement results from self-selected groups of students, who tend to be socioeconomically advantaged relative to the general population in their states.[7]

Further, SAT and ACT scores are questionable measures of the overall achievement of students or of the output of the education system (Selden, 2004). Much of the preparation that students undertake for college-entrance exams occurs outside of the purview of school systems. The content tested on NAEP, in contrast to college-entrance exams, is more aligned to the standards that undergird the curricula of school systems. The standards represented on NAEP tests are reached through discussion and the consensus of state chiefs and subject matter experts (Mullis, 2004).

NAEP also offers the advantage of testing students in 4th and 8th grade, which are grades considered to be critical junctures in the educational trajectory of students (Scott & Ingels, 2007). For instance, 4th grade is when many students are expected to transition from "learning to read" to "reading to learn" (National School Board Association, 2015).

[7] Some states require the SAT or ACT to be administered in high school as part of their accountability systems. However, most states do not.

For these reasons, results from State NAEP serve as a source of data from which researchers and policymakers can gather a sense of how school systems are supporting the learning and achievement of students in a more credible manner than results from college entrance exams. Although results from State NAEP still do not lend themselves to supporting causal inferences, they offer a *more accurate* representation of the outcomes of schooling and, correspondingly, better indications of the efficacy of schools systems in their ability to support learning and achievement than any other assessment program administered across states.

The Case for Studying Subgroups

The history of education in the United States is marked by wide and persistent gaps in achievement between different demographic groups of students. Students from families with higher social standing and/or better economic circumstances tend to perform at higher levels, a phenomenon largely ascribed to an accumulated history of uneven access between groups to economic, social, and cultural resources that are advantageous for succeeding at school (Braun & Kirsch, 2016). For this reason, the efforts of education reformers are often framed by the twin goals of reducing persistent achievement gaps and raising the overall achievement of students.

Part of the solution to reducing achievement gaps is to determine how to best support the learning of historically underperforming subgroups, including students who are low-income, Black, Hispanic, American Indian, learning English as a second language, or have a learning disability. If researchers and policymakers can develop a better sense of which jurisdictions (e.g., states) best support different kinds of students, including historically underserved and underperforming students, then they will be better equipped to recommend and implement policies that support the learning of those students. While NAEP results cannot, of course,

supply researchers and policymakers with a complete answer to this complex question—how to best support the learning of these groups of students—they can offer a partial and helpful answer.

The Case for Full (Complete) Test Samples

For the test sample used in this dissertation, State NAEP results of subgroups on the grade 8 mathematics assessment in 2015, the NAEP program is only able to fully report on eight of eighteen subgroups. In other words, for ten separate reporting groups, the NAEP program was unable to sample a sufficient number of students from at least one state. For example, mean achievement estimates of Black students are available for only thirty-nine of fifty states.

Hence, reporting by State NAEP on the learning and achievement of different students across states is incomplete. A more complete view of subgroup achievement across states would give researchers and policymakers a *fuller* understanding, albeit short of complete understanding, of the extent to which states are supporting the learning of different kinds of students, and, potentially, which sets of policies might best support students from different subgroups.[8]

In addition**,** if NAEP were able to estimate the mean achievement of each subgroup across states, researchers and policymakers would also be able to better discern which states best support student learning *in general*. Full reporting, for instance, would permit the application of Direct Standardization—a statistical technique in which group-specific rates of achievement of a study population are applied to the group-specific distribution of a standard population (Bains, 2009). Applying subgroup-specific estimates of mean achievement to a standard population whose demographic distribution is standardized would provide a basis on which to more fairly

[8] It would also be useful for researchers and policymakers to have a better understanding of the learning and achievement of intersections of subgroups (e.g., Black males). Estimating the achievement of more granular subgroups, however, is beyond the scope of this study. Should one or more of the prediction techniques under examination show promise in the ability to accurately predict mean subgroup achievement, then a logical next line of inquiry would be whether the technique or techniques can also accurately predict the mean achievement of intersections of subgroups.

compare the outputs of states' educational systems, because variation in achievement is strongly associated with the demographic characteristics of students and the demographic distributions of students vary by state.[9]

The implementation of direct standardization requires knowledge of each group-specific rate—in the context of this study, estimates of mean math achievement of subgroups in each state. Subgroup achievement as currently reported by the NAEP program, including the test sample, prevents the ability to conduct this kind of analysis because there are various subgroup estimates of mean math achievement (rates) that are unreported. A distinct value of direct standardization is that it facilitates the identification of "standout" units of analyses (e.g., states). In the context of State NAEP, it permits the researcher to ask— "Holding certain important demographic characteristics of students equal, in which of the 50 states do students have the highest achievement?" Although the answers to this question do not justify causal inferences about the effects of policies on learning, they serve as the basis for enacting potentially useful policy that is grounded in robust research. Alternatively, answers to this question may serve as the basis for conducting follow-up research that might further help clarify which sets of policies might best support student learning.

An additional benefit of applying direct standardization to state NAEP results is that it could help address what can justifiably be perceived as an injustice in how State NAEP results are generally reported and received by the public. In particular, promoting the application of direct standardization to results from State NAEP could help temper the tendency of public

[9] It is important to note that the application of direct standardization does not provide a *completely* fair analytic framework for comparing the outputs of states' educational systems. For instance, in addition to the supports that students receive from their families and school systems, their learning and achievement is influenced to varying degrees by other kinds of resources provided by local and state government, including forms of financial and medical assistance, which vary in availability across states and municipalities.

officials to misinterpret and draw unsupported conclusions from state rankings. Presenting the mean achievement estimates of states as rankings (i.e., league tables), without regard to differences in the demographic composition of students across states, which is long-standing NAEP practice, lends itself to conflating high-achieving students with high-quality school systems. Consequently, the education systems of states like Massachusetts, with relatively large proportions of socially privileged students, are widely celebrated as model education systems. By contrast, this leaves education systems in states such as Alabama and Mississippi, with relatively low proportions of socially privileged students, perpetually cast as deficient.

The Case for Measuring Overall Performance and Performance by Subgroup

The performance of techniques in their ability to accurately predict mean subgroup achievement is evaluated both for states *across* subgroups as well as *by* subgroups of interest. In this dissertation, three techniques are evaluated both in terms of their *general* ability to predict subgroup achievement of states as well as by how well they predict the mean math achievement scores of states for a *particular* subgroup (e.g., states' mean math achievement of Black students). This approach to measuring the performance (predictive accuracy) of techniques permits inquiry into whether accuracy of the techniques varies as a function of subgroup.

The main reason for measuring both *across and by* subgroups is that the sets of findings serve to either corroborate or cast doubt on the efficacy of the techniques under examination. If a technique performs relatively well across subgroups, but performs poorly for one or more subgroups, then doubt should be cast on the technique's ability to estimate subgroup achievement or be useful in practice. On the other hand, if a technique performs well in accurately predicting subgroup achievement both across and within subgroups, then there is

more evidence to suggest that the technique reliably predicts subgroup achievement with a certain desired level of accuracy.

There is also precedence for suspecting that the measures of accuracy may vary by subgroup when comparing NAEP achievement estimates. Hedges and Bandeira de Mello (2013) conducted a study of the validity of NAEP Full Population Estimates (FPEs) by comparing mean achievement results from 2011 special inclusions studies, which involved sampling and testing English learners and students with learning disabilities normally excluded from operational NAEP, to FPEs.[10] One of the study's findings was that the degree of similarity between special inclusion results and FPE results varied by geographic region. Specifically, results were less similar for students from the West and Southeast regions of the United States.

The Case for wMAE and Coverage as Measures of Accuracy

wMAE

The first measure of accuracy used for evaluating the predictive accuracy of techniques in this study, weighted Mean Absolute Error (wMAE), is a variant of the more commonly used statistic, Mean Absolute Error (MAE). MAE, along with measures such as Mean Squared Prediction Error (MSPE; MSE) and Root Mean Squared Error (RMSE), are commonly used to evaluate the performance of predictive models.

Although RMSE and MSPE are more commonly used measures of predictive accuracy (Drakos, 2018), a variant of MAE is chosen for its clearer interpretation and the more proportional nature in which differences between values are factored into the summary statistic (MAE). Regarding clarity, MAE is simply the average absolute difference between values from

[10] Full Population Estimates (FPEs) are estimates of mean achievement based on assumptions about how excluded students (English learners and students with learning disabilities) might have performed on NAEP testing. The National Center for Education Statistics (NCES) has calculated FPEs for states since 2005 and for districts since 2007 (U.S. Department of Education, 2020b).

two variables. RMSE and MSPE, on the other hand, involve an additional arithmetic step, which yields a less intuitive interpretation (Willmott & Matsuura, 2005). Regarding proportionality in treatment of deviations between variables, consider how MSPE is calculated,[11]

$$MSPE = \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n},$$

Whereas MAE is calculated by averaging absolute differences, MSPE is calculated by averaging squared differences. While the intent of these measures is similar, squaring differences results in the undesired effect of giving larger discrepancies disproportionate influence on the value of the statistic (in this case, MSPE), without regard to variability. MAE, on the other hand, preserves the actual magnitude of deviations between variables.

The MAE statistic is expressed as follows,

$$MAE = \frac{\sum_{i=1}^{n}|y_i - x_i|}{n};$$

Where the statistic is equal to the sum of absolute differences over instances of interest between variables *y* and *x*, divided by the number of comparisons (instances of interest) *n* made between variables *y* and *x*. Put differently, MAE is simply the mean absolute difference between variables *y* and *x* over a set number of comparisons (*n*).

The Case for Weighting MAE

In this dissertation, a *weighted* Mean Absolute Error (wMAE) is used in place of *MAE* in order to diminish the relative contribution of absolute differences for which the estimated standard errors associated with the NAEP-reported estimates (y_i) are relatively large. Calculation of wMAE in this study is expressed as,

[11] In the formulas for MSPE and MAE presented on the following page, "y" represents a target value (in this study, a NAEP-reported value), "x" represents values produced from a prediction model (in this study, a technique-produced value) and "n" represents the number of instances in which values "x" are compared to values "y".

$$wMAE = \frac{\sum_{i=1}^{n}(|y_i - x_i| \div SE_{y_i})}{n};$$

As illustrated with the formula, the weighted statistic is calculated in the same manner as *MAE*, but each absolute difference between values from variables *y* and *x* (indexed by *i*) is divided by the estimated standard error associated with y_i.

In this dissertation**,** variable *y* corresponds to NAEP-reported estimates of mean math achievement of states' subgroups (i.e., target values) and their standard errors. By requiring that absolute differences be divided by these standard errors, NAEP-reported estimates of achievement that are calculated with less precision (larger standard errors) have less influence than NAEP estimates calculated with greater precision in determining the value of the accuracy statistic (wMAE). After all, the NAEP reported achievement values are themselves estimates and the estimates calculated with less precision are likely to be less accurate than those with greater precision.

As an instructive example of how this weighting procedure operates, consider two absolute differences of equal magnitude—for example, 4.0, but where the standard error associated with one of the NAEP-reported estimates is twice as large as the other, 2.0 and 1.0. As intended, the deviation associated with the NAEP-reported estimate with a larger standard error will contribute less (2 points) compared to the estimate with a smaller standard error (4.0 points) to the aggregate measure of accuracy (wMAE).

If NAEP-reported estimates of mean achievement were calculated without error, then the average discrepancies between NAEP-reported mean achievement values and technique-produced estimates of mean achievement could be reported as the bias (e.g., through MAE). However, NAEP estimates of mean achievement are subject to both sampling and measurement

error (Reardon, Kalogrides & Ho, 2019). The wMAE in this study can be conceived as a form of measurement error-corrected accuracy statistic.

It is not uncommon to assign more or less influence to particular distances (differences) when computing measures of accuracy (Cleger-Tamayo et al., 2012; Ponomarenko et al., 2010; Gomes et al., 2013). Further, weighting MAE in particular has been employed in previous research. Cleger-Tamayo and colleagues (2012), for instance, demonstrate the use of a weighted measure of MAE to assign more influence to more recent data in a study of recommender systems.

One drawback to weighting MAE *across* subgroups as a function of NAEP-reported precision estimates, is that greater influence in the calculation of the aggregate measure across subgroups is systematically given to differences from demographic subgroups of students that are more prevalent in state populations. Consider for instance that both Hispanic and Asian Pacific Islander are subgroups of interest but that the former is a more populous subgroup across states. Since Hispanic students are more likely to be sampled, they are more likely to have mean math achievement estimates calculated with greater precision (i.e., smaller standard errors). Thus, on average, state-level deviations in the Hispanic subgroup will tend to have more influence (than those in the Asian Pacific Islander subgroup) in the calculation of *overall* wMAE. Nevertheless, this particular drawback is deemed less consequential than not inversely weighting MAE by NAEP-reported standard error estimates. In addition, this concern is somewhat mitigated since the performance of techniques is also compared *within* subgroups.

Coverage

The second measure of accuracy, coverage, is commonly used to calculate the proportion of time prediction intervals contain a true value of interest (Best et al., 2008). In other words,

coverage represents the relative frequency with which a population parameter is contained within the lower and upper bounds of an interval obtained by some statistical procedure.

The coverage statistic used in this study represents a departure from coverage as traditionally calculated since the focal value of interest does not represent the actual parameter of interest but estimates (NAEP-reported estimates of mean math achievement) of the parameter of interest. Further, this measure of coverage reflects the rate at which technique-based estimates of achievement fall within intervals associated with the target value, instead of the rate with which an interval covers the true value of a parameter.

To calculate coverage in this study, let *C(x)* be the number of instances in which technique-produced predicted values of mean math achievement $\{1,\ldots n\}$ fall within 0.2 standard deviations of the corresponding NAEP-reported estimates of mean math achievement. Then, the coverage statistic equals

$$\frac{C(x)}{n},$$

the *proportion* of times that predicted values fall within 0.2 standard deviations of corresponding NAEP-reported estimates of mean achievement of interest (i.e., the estimands). While the implementation of coverage in this study deviates from how the statistic is most commonly calculated, it nonetheless provides a sense of the consistency with which predicted values come close to target values.

Target intervals defined by upper and lower bounds that are 0.2 standard deviations greater and less than target estimates of mean achievement are selected because differences of 0.2 are considered small in size (effect) when standardized mean difference (SMD) is used to measure distances between two means (Cohen, 1988; Lipsey, 2001). In this study, technique-produced estimates of mean achievement represent estimators and NAEP-reported estimates of

mean achievement represent estimands. When the distance between these corresponding values is less than 0.2 standard deviations, the difference between estimator and estimand is considered negligible and the estimator a successful approximation of the estimand.

In this dissertation, the standard deviations from which SMDs are calculated represent the median of the NAEP-reported state-level standard deviations for each subgroup of interest. As such 10 separate standard deviations are used for computing SMDs, one per subgroup of interest. Among commonly used measures of SMD in the research literature, the most similar to the SMD used in this study is Glass's Delta (Glass et al., 1981),

$$\text{Glass's } \Delta = \frac{|M_1 - M_2|}{SD_{control}},$$

which was formulated for measuring standardized effect size in the context of experiments, hence the denominator, $SD_{control}$, the standard deviation of the control group. The numerator is the absolute difference between $M_1$and M_2, the sample means of treatment and control groups. By contrast, the denominator used for calculating SMDs in this study is a standard deviation associated with the estimand (in this study, the median of the NAEP-reported state-level standard deviations for each subgroup of interest), which yields a form of SMD distinct from Glass's Delta, henceforth referred to as *b*, which in general terms is expressed as,

$$b = \frac{|M_{estimator_{ij}} - M_{estimand_{ij}}|}{SD_{estimand_I}},$$

where the difference between estimator and estimand for subgroup *i* in state *j* is divided by the standard deviation of subgroup *I*. An aggregate (i.e., the median) standard deviation associated with the estimand, $M_{estimand_{ij}}$, is used for the denominator, $SD_{estimand_I}$, since NAEP-reported standard deviation estimates for individual state-subgroup pairs are unstable. While the majority of these standard deviation estimates range between values of 30 and 35, some are greater than 45 and at least one is as large as 51 (U.S. Department of Education, n.d.). Thus, using the

reported standard deviations of estimands $SD_{estimand_{ij}}$ would unduly favor the performance of technique-based estimates for estimands with relatively large standard deviations. The target intervals would be inappropriately large.

Other common forms of SMD, including Cohen's d (Cohen, 1988) and Hedges' g (Hedges & Olkin, 1985), use a denominator that represents a pooled standard deviation—a measure that represents a weighted average of the standard deviations of treatment and control samples. Using a pooled standard deviation to compute this study's SMD (b) is unsuitable for numerous reasons. First, the technique-based estimates are calculated from multiple samples. Second, while variance estimates can be computed for technique-based estimates of mean math achievement during the analysis phase of this dissertation (chapter 4), standard deviations are required for a simulation study that must precede the analysis. Results from the simulation inquiry are used to determine the rate at which estimators can reasonably be *expected* to fall within target intervals, which informs an apriori specification of coverage-rate (i.e., hit-rate) criteria for determining whether a technique could be recommended for use in practice.

The Case for Using Competing Approaches to Predict Mean Subgroup Achievement

Three analytic techniques are used and evaluated to predict mean subgroup achievement on State NAEP's grade 8 mathematics assessment from 2015: MICE, FH and FLEX CS. The rationale for using multiple analytic methods for the same purpose reflects a belief that scientific research is a continual process of reasoning supported by an interplay of methods, theories and findings (Shavelson & Towne, 2002). In this inquiry, it is assumed that relative performance of each technique—in terms of its ability to accurately predict subgroup achievement—is a function, in part, of the type of variables and algorithms that underpin each analysis. Further, it is assumed that each set of findings resulting from the three techniques offer the opportunity to

glean important information that is potentially useful for follow-up research. It is conceivable, for instance, that none of the three techniques perform particularly well, but the sets of results reveal information that leads to another promising approach for predicting subgroup achievement.

The Case for Imputation

Imputation refers to a family of statistical techniques for replacing missing data with substitute data values. While different rationales underpin the mechanisms by which different imputation techniques operate, all attempt to replace missing values with values that might reasonably have been expected. Imputation thus can be conceived as a type of prediction technique.

The use of imputation in this study strays from its more common application. In order to evaluate the predictive accuracy of the imputation technique used in this study, target values (i.e., NAEP-reported estimates) are successively withheld prior to each imputation to permit a comparison between imputed (predicted) values and the target values (NAEP-reported estimates). The target value is then returned to the dataset before withholding a different target value, and so on. This adaptation serves the evaluative nature of this study. The evaluation of predictive accuracy of techniques under examination is based on comparisons between predicted and observed (target) values. By contrast, imputation is not applied to a real world missing data problem in this study (more detail on this "withholding" process is offered in chapter 3).

The Case for Multiple Imputation

Among the myriad of imputation techniques for handling missing data, *modern* missing data analysis techniques are generally preferred over traditional techniques such as mean substitution since modern methods more effectively preserve the relationships between variables

from the original dataset (Schafer & Graham, 2002; Johnson & Young, 2011). Modern imputation techniques generally involve variants of Multiple Imputation (MI), although some researchers consider single-imputation with the Expectation Maximization (EM) algorithm to be a form of modern missing data analysis as well (Graham, 2009; Johnson & Young, 2011).

MI techniques produce, by definition, multiple imputed datasets and account for the uncertainty introduced by the presence of missing values through a series of random draws from predictive distributions (Johnson & Young, 2011). Multiple draws permit the quantification of uncertainty due to missing data and, because of this treatment of uncertainty, many researchers consider MI to be the gold standard method for handling missing data in statistical research. Single-value imputation techniques, by contrast, offer only one substitute value per missing value and thus do not lend themselves to the quantification of uncertainty.

MI is particularly fitting for this study because it has been shown that the approach performs well in small samples (Graham, 2009), including samples smaller than 50 observations (Barnes, Lindborg & Seaman, 2006)—the number of observations in the test sample. Using simulated data, Barnes and colleagues (2006) demonstrate, for instance, that certain forms of MI produce relatively accurate estimates of missing data with sample sizes as small as 20.

Previous research also indicates that MI performs well when there is as much as fifty percent missing data in variables to be imputed, with missing data presumed to be missing-at-random (MAR) (Graham & Schafer, 1999). While most of the subgroups of interest in this study—which play role of variables to be imputed—do not exceed more than fifty percent missing data, half (5 of 10) have relatively high proportions of missing data (i.e., > 20%). These include variables representing estimates of mean math achievement of English learners as well as students who identify as Black, Asian or Pacific Islander, American Indian or Alaskan Native,

and multiracial. Further, research suggests that certain forms of MI produce accurate estimates, based on measures of coverage and standardized bias, when there is as much as forty percent missing data in the overall test sample (Barnes, Lindborg & Seaman, 2006).

The test sample used in this study is represented by a fifty-by-eighteen data matrix, in which rows correspond to the fifty American states and columns to State NAEP's reporting groups (subgroups). Overall, the test sample has about fourteen percent (13.8%) missing data.[12] Given MI's robustness to high proportions of missing data, as well as its ability to produce accurate estimates with small samples, the technique is a safe and credible choice for prediction in this study.

The Case for MICE

The form of MI used in this study is Multivariate Imputation by Chained Equations (MICE). This form of MI is referred to as a fully conditional specification (FCS) approach to imputation, and sometimes as sequential regression multiple imputation (Azur et al., 2011), since the imputation model is specified on a variable-by-variable basis by a set of conditional distributions produced from regressions, one for each incomplete variable. Once values for one variable are imputed, the algorithm governing the MICE procedure imputes the values of a new variable in a sequence, with predictor variables that are specified by the researcher at the outset of the procedure. Uncertainty is incorporated into each step of imputation through simple random draws from the conditional distributions. These steps are repeated a pre-determined number *m*

[12] It should be noted that the missing data mechanism (MCAR vs MAR vs MNAR) in this study is well understood. Values are missing from the dataset because they represent students from subgroups across states that the NAEP program was unable to sufficiently sample. It is not assumed that the missing data mechanism introduces bias, as might be expected, for instance, if the values were missing from the dataset because they represent unusually low or high levels of achievement.

times and results in *m* separate imputed data sets (greater detail regarding this imputation process is offered in chapter 3).

The variables used for prediction are referred to as auxiliary variables. These variables are not of primary interest but enter the imputation models as they are related to incomplete variables (the variables of interest in this study) and support accurate imputation (prediction) of missing values. MICE is chosen over other forms of MI, such as Multiple Imputation with the Expectation Maximization (MI EM) algorithm for MICE's flexibility in model specification, which is a useful feature for reducing bias in missing value estimates (Graham, 2009; Johnson & Young, 2011; Collins, Schafer & Kam, 2001). MICE permits the researcher to select which variables from the original dataset will be used to guide imputation of data variables with missing values on a variable-by-variable basis. As such, the researcher is able to select only the variables from the original dataset that are most strongly associated with each variable with missing data to play the role of predictor variables.[13]

The application of MICE is particularly useful for this study since, unlike other MI procedures, MICE facilitates separate model specification for each variable to be imputed and the variables to be imputed in this study are highly correlated with some variables from the test sample, but are unrelated with others. For instance, variables representing the mean math achievement of parental level of education subgroups are generally highly correlated with one another, as might be expected, and serve as promising predictors for one another. Meanwhile a pattern of strong correlations between these parental level of education variables and other variables does not appear. For instance, there are weak correlations between most parental level

[13] As can be gleaned from the corresponding paragraph, separate imputation models are specified for each incomplete variable. Thus, there is no standard model used for imputing the incomplete variables (more detail on the specification of imputation models is provided in chapter 3.

of education and race/ethnicity subgroups, and hence less reason to believe they would serve as effective predictors for one another.

The Case for Small Area Estimation (SAE)

Small Area Estimation (SAE) is an analytic framework for estimating the parameters of subpopulations when the samples available for producing direct estimates (design-based estimates) of the subpopulation parameters are insufficiently large, which in turn does not permit for calculation of estimates with sufficient precision (Ghosh & Rao, 1994). Using an SAE technique represents a logical approach for dealing with the problem that this study addresses. The NAEP program does not always report the mean achievement of minority populations ("small areas") since the program only reports subgroup results if it samples at least 62 students who identify with the subgroup (Elliott & Phillips, 2004; Chromy, Finker & Horvitz, 2004). This dissertation attempts to determine whether a statistical technique can justifiably be applied to subgroup achievement estimation on State NAEP when direct estimation is impermissible because of insufficient sampling.

Estimation in SAE involves borrowing information from administrative data to improve direct estimates of interest. To improve direct estimates in the SAE framework typically means combining direct estimates with synthetic (model-based) estimates. Because the measures of accuracy in this study, by which the predictive performance of techniques are evaluated, require use of target values that *are not* estimated with insufficiently small samples, an adaptation of SAE is implemented whereby small random samples from larger samples of available data are used as direct estimates. Hence, this study simulates the experience of failing to obtain at least 62 students from a subgroup (more detail on this adaptation is offered in chapter 3).

SAE is widely used by agencies within the U.S. federal government. As of 2016, at least eight separate U.S. agencies were implementing an SAE program—including the Census Bureau, Bureau of Labor Statistics, National Center for Health Statistics, Agency for Healthcare Research and Quality, National Cancer Institute, Bureau of Justice Statistics, Department of Agriculture, and National Center for Education Statistics (Czajka, 2016). An often-cited example of a successful application of SAE is the U.S. Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program (Beresovsky & Hsiao, 2014; National Research Council, 2000). The program provides income and poverty estimates for U.S. counties and fulfills a legislative mandate to produce yearly estimates of children living in poverty within local jurisdictions across the United States, which serves to guide the allocation of federal funds across local jurisdictions.

A persistent challenge in SAE is the identification of high quality administrative data for computing model-based estimates (Czajka, 2016; Rao, 2012). Model-based estimates produced from administrative data primarily serve to render small area estimates more precise. However, administrative data that are not particularly highly correlated with the phenomenon of interest can introduce unwelcome bias in the resulting estimates. This dissertation draws on administrative data from various sources that are highly correlated with the data variable of interest, mathematics achievement (e.g., socioeconomic data from the U.S. Census Bureau).

The Case for the Fay-Herriot (FH) Model

SAE models are generally classified or described by the type of administrative data that are used to compute synthetic (i.e., model-based) estimates, which are then combined with direct estimates to form precision-weighted estimates of the area-level parameter of interest. Two broadly defined SAE models are the *area-level* and *unit-leve*l models. Area-level models involve

use of administrative data from the same level of aggregation as the small area parameter of interest to form synthetic estimates, while unit-level models involve using administrative data measured and collected from lower levels of aggregation.

As an instructive example of an area-level model, consider how the U.S. Census Bureau estimates poverty rates of certain U.S. counties with small populations in its SAIPE program (Small Area Income and Poverty Estimates). Because the Census Bureau only directly samples a limited number of households in certain small counties (which results in imprecise estimates), the Bureau uses county-level data that are correlated with county-level poverty rates from the Internal Revenue Service (IRS) and the Supplemental Nutrition Assistance Program (SNAP) in a linear regression model to produce regression-synthetic (model-based) estimates of county poverty rates. These synthetic estimates are then combined with direct estimates to form precision-weighted parameter estimates for the "small areas." On the other hand, if the Census Bureau were to use administrative data form the IRS and SNAP that are measured at lower levels of aggregations (e.g., municipalities, households) to produce synthetic estimates, they would be using a unit-level model.

In this study, a commonly used area-level model known as the Fay-Herriot (FH) model (the same model used by the SAIPE program), sometimes referred to as "area level random effects model" (Pfefferman, 2002), is used to predict mean subgroup achievement on State NAEP. The parameter estimate produced by the FH model is referred to as an Empirical Best Linear Unbiased Predictor (EBLUP), which is a precision-weighted combination of the direct and synthetic estimates (Molina & Marhuenda, 2015). Calculation of the EBLUP can be expressed in the general form,

$$\hat{\delta}_d^{EBLUP} = \hat{\gamma}_d \hat{\delta}_d^{DIR} + (1-\hat{\gamma}_d) x_d^{\mathrm{T}} \hat{\beta},$$

Where $\hat{\delta}_d^{DIR}$ is the direct estimate of the subpopulation (small area) parameter of interest and $x_d^{\mathrm{T}}\hat{\beta}$ is a regression-based (model-based) estimate of the subpopulation parameter of interest, and $\hat{\gamma}_d$ represents the proportion of total error variance attributable to the regression estimator (i.e., relative precision of the direct estimate). As such, the greater the uncertainty of the regression estimate relative to that of the direct estimate, the more the EBLUP is shifted toward the direct estimate. For this reason, the EBLUP is also known as a "shrinkage" estimator, as the regression estimate "shrinks" back toward the direct estimate to a degree commensurate with the precision with which the direct estimate is calculated relative to the calculated precision of the regression estimate (greater detail on the specific application of the FH model in this study is provided in chapter 3).

Although comparative evaluations of competing SAE approaches are limited in the research literature (Czajka, 2016), at least two studies indicate better predictive accuracy of area-level models compared to unit-level models (Gomez-Rubio et al., 2010; Best et. al, 2008). Using simulated data, Gomez-Rubio and colleagues (2010) compare the predictive accuracy of various area- and unit-level models with measures of Mean Absolute Relative Bias (MARB) and find that the area-level models consistently produce estimates with smaller MARB (less bias). Similarly, albeit with measures of Average Empirical Mean Square Error (AEMSE) and by randomly sampling from real data on household income in Sweden, Best and colleagues (2008) find that area-level models perform better in terms of predictive accuracy. As a possible reason for the greater predictive performance, Best and colleagues note that area-level models may be more robust to the presence of anomalous observations at the unit-level given area-level models fit aggregate data. In practice, the choice between using an area- or unit-level model is often dictated by the data that are available (Gomez-Rubio et al., 2010).

Besides research evidence to support the use of area-level over unit-level models for producing accurate parameter estimates, the use of FH, an area-level model, is suitable for this study as there are various rich sources of publicly-available administrative data measured at the area-level (state-level), which are highly related to measures of academic achievement. Examples of these sources of data include the American Community Survey (ACS) and the Common Core of Data (CCD). The availability of such data permits the construction of regression-models that should produce reasonably accurate estimates of mean subgroup achievement across states. These area-level data are combined with direct estimates of mean subgroup achievement, which are averaged from restricted-use student-level data from the National Center for Education Statistics (NCES).

The Case for the Selected Administrative Data

Accurate estimation of synthetic-based values is contingent on use of good administrative data (Rao, 2012). That is, sets of variables that can accurately predict true parameter values. For each subgroup of interest in this study, for which NAEP-reported mean estimates of subgroup achievement across states play the role of response variable in regression-based estimates, administrative data that have an empirical and theoretical relationship with measures of mean achievement are used as predictor variables.

Predictors for Achievement of Parental Level of Education Subgroups

To calculate synthetic-regression estimates of mean math achievement for the first four subgroups of interest, which represent students whose parents have different levels of educational attainment, state-level variables representing four separate factors are used as predictor variables: students' race and ethnicity, the economic circumstances of students' families, the English proficiency of students, and the quality of schooling that students

experience. Race and ethnicity are represented by a dichotomized variable operationalized as each state's overall percent of grade 8 students who identify as Black, Hispanic, American Indian, or Alaskan Native (*%B-H-AIAN*),—historically marginalized and oppressed subgroups of students. The economic circumstances of students' families is represented by a variable that reflects a composite measure of a state's median household income and wealth (Family Economic Resources; *FER*). English proficiency is represented by a variable operationalized as each state's percent of students identified as English learners (*%EL*). School quality is represented by a variable based on the scores, reported annually by Education Week (2015), related to each state's effort to improve public education (*SQI*) – an indicator of school quality measured on a continuous scale ranging from 0 to 100.

Figure 2.1: Regression models for computing state-level synthetic estimates of mean math achievement of students from different parental level of education subgroups

$$\hat{Y}_{\overline{NHS}_{math}} = \hat{\beta}_{intercept} + \hat{\beta}_1(\%\text{B-H-AIAN}) + \hat{\beta}_2(\text{FER}) + \hat{\beta}_3(\%\text{EL}) + \hat{\beta}_4(\text{SQI})$$
$$\hat{Y}_{\overline{HS}_{math}} = \hat{\beta}_{intercept} + \hat{\beta}_5\ (\%\text{B-H-AIAN}) + \hat{\beta}_6(\text{FER}) + \hat{\beta}_7(\%\text{EL}) + \hat{\beta}_8(\text{SQI})$$
$$\hat{Y}_{SBA_{math}} = \hat{\beta}_{intercept} + \hat{\beta}_9(\%\text{B-H-AIAN}) + \hat{\beta}_{10}(\text{FER}) + \hat{\beta}_{11}(\%\text{EL}) + \hat{\beta}_{12}(\text{SQI})$$
$$\hat{Y}_{\overline{BA}_{math}} = \hat{\beta}_{intercept} + \hat{\beta}_{13}\ (\%\text{B-H-AIAN}) + \hat{\beta}_{14}(\text{FER}) + \hat{\beta}_{15}(\%\text{EL}) + \hat{\beta}_{16}(\text{SQI})$$

Race & Ethnicity Predictor (%B-H-AIAN)

Race and ethnicity in this study refers to social categories related to ancestral origin, to which residents of the United States self-identify. Using guidelines from the U.S. Department of Education (2007), ethnicity here is used to identify students as "Hispanic or Latino," regardless students' race. Race, on the other hand, is used to categorize students into four separate groups: American Indian or Alaskan Native, Asian or Pacific Islander, Black or African American, and White. In addition, the U.S. Department of Education provides students the opportunity to identify with "two or more races." Though the constructs of race and ethnicity are complicated,

for practical reasons in educational research, data variables that represent race and ethnicity are often used to capture whether students identify with a historically and/or contemporaneously higher- or lower-achieving racial or ethnic group. In the context of American schooling, because of uneven access to economic, social and cultural resources advantageous to academic success—largely shaped by an enduring history of racial oppression—students who identify as Black, Hispanic and American Indian tend to underperform other racial and ethnic groups on most measures of academic achievement.

To compute synthetic-regression estimates of mean math achievement for subgroups representing students whose parents attained different levels of education, a composite race and ethnicity predictor variable is used that represents multiple subgroups. This variable, *%B-H-AIAN,* includes values that reflect the combined percent of students by state who identify as Black, Hispanic, American Indian, or Alaskan Native—groups of students that have historically underperformed on measures of academic achievement when compared to other subgroups.

Rationale for %B-H-AIAN Predictor. Though differences in socioeconomic status (SES) accounts for substantial variation in achievement differences between students (Coleman et al., 1966; Reardon, 2011), research continues to document lagging performance of certain historically marginalized and oppressed racial and ethnic minority groups of students on measures of academic achievement, even after holding indicators of SES constant (U.S. Department of Education, 2001; Ogbu, 2003). Using NAEP mathematics data from 2011 and Common Core of Data from 2010-11, for instance, Bohrnstedt and colleagues (2015) find a nationwide within-school Black-White achievement gap after accounting for students' SES, teacher characteristics, and school characteristics. In her 2006 presidential address to members of the American Educational Research Association, Gloria Ladson-Billings, drew attention to this

phenomenon, sharing that—"even when we compare African Americans and Latinos with incomes comparable to those of Whites, there is still an achievement gap as measured by standardized testing" (Gladson-Billings, 2006, p. 4).

Several theories have been put forth to account for the phenomenon by which certain racial and ethnic minority students underperform compared to socioeconomically similar peers. Ogbu and Simmons (1998) make sense of the issue through cultural-ecological theory, or as they sometimes call it a *cultural-ecological theory of academic disengagement*. Ogbu (2003) argues that the legacy of racial discrimination in the United States has engendered in African-Americans a disaffection toward schooling and reluctance to believe the education system can offer them the opportunity to experience academic success and social mobility. While Ogbu's research on academic disengagement has drawn criticism by other scholars, Ogbu's work has nonetheless provoked helpful debate around why certain racial and ethnic minority students lag behind socioeconomic peers on measures of academic achievement (Foley, 2004).

Another theory put forth is that of "stereotype threat," made popular by the work of Steele, Aronson and Spencer (Steele & Aronson, 1995; Steele, Spencer & Aronson, 2002). The researchers posit that even passing reminders that someone belongs to a certain social group, which is stereotyped as inferior, can hurt an individual's test performance. Steele and Aronson (1995), for instance, using items from the Graduate Record Exam (GRE), found that making Black test takers aware of the stereotype that Black people are less academically capable had the effect of depressing their scores.

An additional body of research indicates that the behavior of educators adversely impacts the achievement of minority students (Carter, 2008; Warikoo et al., 2016). Carter (2008) argues that when students of color are made hyper-visible or ignored in the classroom because of their

race, they sometimes cope in ways that lead to academic disengagement. Others who underscore the behavior of educators as a contributing factor to underperformance of minority students, point to a social mismatch between students and educators. Gay (2002) explains that the greater difference there is between students' cultural, racial, and ethnic characteristics, and the normative standards of schools, the greater are the chances their school achievement will be compromised by low or negative teacher expectations. Low teacher expectations are widely understood to negatively influence student achievement (McKown & Weinstein, 2008; Workman, 2012). Gay (2002) explains that the cultural experiences of students of color lead them to be less attuned to the normative standards of schools, which results in unfair teacher attitudes, expectations and actions toward racial and ethnic minority students.

It should be carefully noted that none of these theories suggest that differences in the achievement of racial and ethnic groups are related to differences in innate abilities across groups. Instead, the theories suggest that a legacy of oppression and discrimination have cultivated perceptions in teachers and students that have unfavorable influence on the academic outcomes of certain racial groups of students. Further, it should be noted that the *%B-H-AIAN* variable is used to predict achievement estimates in this study because membership in the *B-H-AIAN* subgroups is *associated* with measures of academic achievement, not because membership causes or results in differences in academic achievement. This is an important last point, as social scientists have a penchant for carelessly using language that evokes the racist idea that race causes differences in academic achievement (Zuberi, 2000).

Family Economic Resources (FER)

Family Economic Resources (FER), in this study, is a measure meant to represent the combined income and wealth of parents or guardians of students who share the same primary

household or place of residence. Income includes salaries and wages, retirement income, government assistance, and investment gains. Family income is similar to the more oft cited statistic, "household income" though the two are different in that household income encompasses the income of all people sharing a common primary place of residence (not only parents or guardians). Family wealth, on the other hand, refers to net worth—the summed value of all of a family's assets (e.g., home, savings in bank), minus liabilities that the family might owe (e.g., credit card debt). Family wealth is more difficult to measure than income, because of limited availability of data. Thus, family wealth, including its influence on student achievement, is less frequently used in models applied in social science research than family income. Nevertheless, research indicates that parents draw on both income and overall wealth to support the academic learning and achievement of their children.

Rationale for FER Predictor. Often cited alongside parental level of education as a leading social factor accounting for differences in achievement of students is the economic circumstances of students' families (Dahl & Lochner, 2012; Manna 2013). A large body of research points to a strong positive relationship between economic resources of students' families and the academic achievement of students (Coleman et al., 1966; Bowles & Gintis, 1976; Reardon, 2011; Braun, 2016).

Affluent parents advantage their children by drawing from their financial resources in various manners. These parents, for instance, can ensure that their children attend well-funded schools with children from other affluent families by settling into homes within school districts where the price of homes render the prospect of settling into the same school district cost-prohibitive for less affluent families. In addition, affluent families are better able to absorb the costs associated with sending their children to private schools (e.g., tuition) compared to less

affluent families. Meanwhile, affluent families can also bestow academic advantage upon their children by providing them private tutoring or academically enriching opportunities after-school or during breaks in the academic calendar (Bourdieu, 1986; Bowles & Gintis, 1976).

Besides these rather obvious ways by which affluent families can leverage financial resources to support their children's academic success, money—or the lack thereof—influences the learning and achievement of students in several less obvious ways. Middle- to upper-income families generally provide healthy childhood environments, including a regular supply of nutritious food, housing stability and feelings of security, quick medical or dental attention when needed, high-quality childcare, and access to educational resources such as books and computers. In contrast to middle- and upper-income families, low income families' residential options are more limited, often restricting them to live in neighborhoods with high concentrations of other low-income families, where economic opportunities and prospects of upward mobility are scarce. Children who grow up in concentrations of poverty are disproportionately vulnerable to a variety of health risks, including otitis media (ear infections), asthma, lead poisoning, and mercury poisoning, all of which weigh negatively on learning and achievement (Braun, 2016; Berliner, 2013).

English Learner Predictor (%EL)

English learners (ELs) are students who are unable to communicate fluently in English or learn effectively in English. ELs come from non-English speaking homes and typically receive specialized or modified instruction to accommodate their English language limitations. The predictor variable, *%EL*, represents the percent of students within states identified as English learners.

Rationale for %EL Predictor. On the 2015 grade 8 NAEP-mathematics assessment, the achievement of 8th graders (nationwide) who were not identified as English learners was 1.03 standard deviations greater than the achievement of those who were identified as English learners. On the corresponding reading assessment, 8th graders that were not identified as English learners scored 1.29 standard deviations greater than those identifying as English learners.

Besides the obvious language barrier that hinders the achievement of English learners, research points to different features of schooling that further undermine opportunities for English learners to achieve, including widespread lack of educator preparation or resources to support the learning needs of English learners (McGraner & Saenz, 2009; American Psychological Association, 2012). For instance, a growing body of literature highlights a lack of linguistically responsive pedagogy and dual language instruction in American schools—both of which are recommended instructional approaches to support the learning of English learners (Lucas, Villegas & Freedson-Gonzalez, 2008; American Psychological Association, 2012; Gandara & Rumberger, 2009).

Dual language instruction, sometimes referred to as bilingual education, is critical to the learning of language minority students. Conclusions from five separate meta-analyses confirm that children who receive instruction in their native language have higher rates of academic achievement, even when the markers of achievement are in English, compared to their peers who receive less instruction in their native language (American Psychological Association, 2012). One theory to account for the phenomenon by which dual language students outperform English-immersion students holds that first bolstering literacy in one's native language helps English learners more quickly grasp the syntax and rules of a second language.

School Quality Index (SQI)

While out-of-school factors account for a majority of variation in student achievement (Coleman et al., 1966; Egalite, 2016), differences in the quality of schooling that students experience also account for differences in achievement between students (O'Day & Smith, 2016). In this study, in addition to sociodemographic measures, a measure of school quality based on ratings calculated and reported by Education Week (2015) is used to model the relationship between school quality across states and mean math achievement to calculate regression-synthetic estimates. The variable, *SQI*, is measured on a continuous scale with scores ranging 0.0 to 100.0, and reflects the average of states' "Chance for Success" and "School Finance" ratings. The Chance for Success rating is meant to capture lifelong learning opportunities for students, beginning with early childhood, and progressing through K-12 education into adulthood. The School Finance rating is based on school spending patterns as well as how education dollars are distributed across each state (Education Week Research Center, 2015).[14]

Rationale for SQI Predictor. While research makes clear that factors aside from the quality of schooling influence differences in NAEP scores across states, research also suggests that, after controlling for factors beyond the control of state systems, significant between-state variation in performance remains (Loveless, 2013; Carnoy, Garcia & Khavenson, 2015; Chingos, 2015). While this remaining variation between states can be, in part, ascribed to differences in the quality of district- and school-level systems, there is also good reason to suspect differences in the quality of state-level systems contribute to difference in learning and achievement as well.

[14] Hawaii is a single-district jurisdiction. As a result, it is not possible to calculate "financial equity," a subcomponent of Education Week's "school finance" measure, which is defined as the equitable distribution of funding across districts within a state. As a result, Hawaii's "school finance" rating, one of the two ratings that are averaged to calculate each state's School Quality Index (SQI) score is measured differently than other states.

States directly influence several important components of schooling, including components measured by the *SQI* predictor variable. The measures that form the *SQI* predictor variable represent efforts and systems that states put in place to support the learning and achievement of students, including decisions around learning standards and curriculum, certification and licensing requirements to teach, and how to generate and allocate a substantial amount of school funding.

Predictors for Achievement of Black Students

To calculate regression-based estimates of mean math achievement for the fifth subgroup of interest, students identifying as Black, two of the state-level variables used for predicting the parental level of education subgroups are again used. These are the data variables representing factors related to the economic circumstances of students' families (*FER* variable) and quality of schooling (*SQI* variable).

In addition, predictor variables representing factors related to parental level of education and Black ethnicity are used to predict the mean math achievement of students identifying as Black. Parental level of education is operationalized as the percent of adults by state that have earned a bachelor's or more advanced degree (*%BA*). Black ethnicity is used to distinguish Black students who identify as African-American from Black students who do not identify as African American. The predictor variable for this factor is operationalized as the percent of the Black population by state who identify as African-American (*%AA*).

Figure 2.2: Regression model for computing state-level synthetic estimates of mean math achievement of Black students

$$\widehat{\boldsymbol{Y}}_{\overline{\boldsymbol{B}}_{\boldsymbol{math}}} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{o}} + \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}(\text{FER}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}(\text{SQI}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{3}}(\%\text{BA}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{4}}(\%\text{AA})$$

Parental Level of Education (%BA)

Parental level of education is defined by the National Center for Education Statistics (U.S. Department of Education, 2016) as the highest level of education of either parent or guardian. Thus, this construct is typically measured in an ordinal manner, whereby variable values are discrete and hierarchical (e.g., high school degree vs. associate's degree; bachelor's degree vs. advanced degree). For the purpose of computing regression estimates in this study, however, parental level of education is represented by a dichotomized variable that reflects the percent of adults by state that have earned a bachelor's or more advanced degree (*%BA*).

Rationale for %BA Predictor. Research on the relationship between social background factors and achievement frequently points to parental level of education as one of the factors with the strongest relationships to achievement (Coleman et al., 1966; Dubow, Boxer & Huesmann, 2009; Reardon, 2011; Manna, 2013, Egalite, 2016). As Hanushek and colleagues (2013) explain, many studies indicate that educational attainments of the mother and father are likely more influential in test performance and life outcomes than any other single variable, including the student's race, household income, or family structure (one- or two-parent home).

One of the more prominent explanations for the disparity in educational achievement between students of higher and lower levels of parental education attainment is Pierre Bourdieu's theory of cultural and social reproduction (Bourdieu & Passeron, 1977; Bourdieu, 1986). The theory holds that social actors consciously and subconsciously shape and exploit institutional structures that permit them to preserve the prevailing stratified social order across generations (Edgerton, Peter & Roberts, 2014). Privileged parents, the theory follows, can use social connections (i.e., social capital) and cultural knowledge (i.e., cultural capital) to facilitate their children's attainment of social advantages.

Educated parents confer academic advantage on their children by transmitting cultural values and forms of behavior, such as attitudes toward schooling and patterns of speech, that are favorable for succeeding in school (Bernstein, 2003; Lareau, 2002). Bernstein (2003), for instance, explains that schools are built on the middle- and upper-classes' *elaborated speech code*, and that teachers judge students who do not use the middle and upper classes' form of speech to be less intelligent, a judgement that is then both explicitly and implicitly communicated to students. Lareau (2002) argues that middle- to upper-class parents rear their children in a manner that she describes as *concerted cultivation*, which is characterized, in part, by consciously fostering the development of language favorable for navigating social institutions.

Integral to Bourdieu's theory of social and cultural reproduction is the concept of *habitus*—an individual's way of thinking, perceptions and dispositions, which are informed by present and past experiences. A child's habitus, largely rooted in familial socialization, shapes the student's outlook on the world, including perceptions of what is possible and preferable for someone from their social position and upbringing. The circumstances in which individuals undergo socialization impact the way they conceive of different roles, including the role of student (Pallas, 1993). Students whose parents and adult role models did not complete high school, for instance, are less likely than other students to view their role of student as one involving academic success. On the other hand, the children of college-educated parents are more likely to develop worldviews favorable for attending college themselves.

Black Ethnicity (%AA)

Research on achievement differences between groups of Black students is uncommon. The limited amount of research is due, at least in part, to the manner by which students are asked to identify themselves, including on NAEP assessments. Black students do not typically have the

opportunity to provide more details about their identity. Some research, however, points to differences in academic outcomes between Black students whose ancestors were forced to the United States many generations ago as slaves compared to students who are immigrants or whose parents are immigrants (Anderson, 2015). The *%AA* variable in this study represents the percent of the Black population in each state who identify as African American, and it is used in an attempt to improve synthetic-regressions estimates of mean math achievement of Black students across states.

Rationale for the %AA Predictor. A 2015 study from the Pew Research Center on the characteristics of Black immigrant populations in the United States points to substantial differences between Black immigrants and Black Americans on measures of academic and occupational achievement (Anderson, 2015). The study, for instance, finds that Black immigrants are three times more likely to hold a college degree. Considering that parental level of education is strongly associated with academic achievement, first-generation Black students presumably fare better on measures of academic achievement than the wider Black population, though a strong research base to support these quantitative differences in achievement between Black populations does not yet exist.

On the other hand, there is a strand of anthropological research that suggests there are systemic differences in the academic orientations and achievement of African American students and other Black students in the United States, differences shaped by distinct minority experiences (Ogbu & Simmons, 1998; Ogbu, 2003). Ogbu wrote of the distinct experiences and socialization of "involuntary" and "voluntary" minorities. The former generally refers to African Americans, Mexican Americans and American Indians, groups whose families have lived in the United States for many generations. The latter generally refers to children of immigrants, who

are racial and ethnic minorities in the United States, but whose families *chose* to come to the United States. For Ogbu, the American experience of voluntary minorities is characteristically one of assimilation and optimism regarding their ability to experience academic achievement and upward mobility. On the other hand, involuntary minorities, including African Americans, are generally encultured to believe that schooling will not help them experience upward mobility (Foley, 2004). "Involuntary" minorities are keenly aware of U.S. institutions' enduring role in discriminating against members of their race or ethnicity and are more reluctant to believe schooling can serve as a vehicle for experiencing upward mobility.

Predictors for Achievement of Hispanic Students

To calculate regression-based estimates of mean math achievement for the sixth subgroup of interest, students identifying as Hispanic, state-level variables representing factors related to parental level of education (%BA), the economic circumstances of students' families (FER), English language proficiency of students (%EL), and the quality of schooling (SQI) are again used as predictor variables. In addition, a predictor variable representing *Hispanic origin* is used. This last predictor variable is operationalized as the percent of the Hispanic population by state of Mexican descent (*%MX*).

Figure 2.3: Regression model for computing state-level synthetic estimates of mean math achievement of Hispanic students

$$\widehat{\boldsymbol{Y}}_{\bar{\boldsymbol{H}}_{\boldsymbol{math}}} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{o}} + \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}(\%\text{BA}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}(\text{FER}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{3}}(\%\text{EL}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{4}}(\text{SQI}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{5}}(\%\text{MX})$$

Hispanic Origin (%MX)

In broad terms, Hispanic (or "Latino") in the United States refers to persons who descend from Spanish-speaking populations and cultures. The U.S. Census collects information on the ancestral countries and regions of Hispanics and designates six "origin types"—Mexican, Puerto

Rican, Cuban, Central American, South American, and Other Hispanic (U.S. Census Bureau, 2016). The predictor variable (*%MX*), used to model the relationship between Hispanic origin and achievement, represents the percent of Hispanics by state whose descendants arrived in the United States from Mexico. This origin type alone accounts for about sixty percent of Hispanics residing in the United States (U.S. Census Bureau, 2016).

Rationale for %MX Predictor. Despite their grouping as a single ethnicity, Hispanics residing in the United States are not a monolithic population. Groups of Hispanics arrived in the U.S. in several waves of migration, from different regions of Latin America, and for various reasons. While the majority of some Hispanic origin groups migrated to the U.S. within the past few generations, others have lived in what is today the Southwestern United States since the early 19th century. The Hispanic population that has long resided in the American Southwest mainly identifies as Mexican American.

Hispanics of Mexican descent, the largest group of Hispanics residing in the U.S., have lower average levels of academic attainment than the other five origin types designated by the Census. In 2016, about twelve percent of Hispanic adults in the U.S. identifying as Mexican held a bachelor's or more advanced degree. By contrast, over twenty percent of Hispanics by origin type other than Mexican held a bachelor's or more advanced degree.

In a sense, the difference in academic attainment of Hispanics of Mexican origin and other Hispanics supports Ogbu's (2003) theory around the academic orientations of voluntary and involuntary minorities. The families of many Mexican American students have been in the U.S. for several generations. By contrast, the new immigrant experience is more characteristic of other Hispanic groups. From the theoretical perspective developed by Ogbu, these more recent

immigrants are more likely to harbor views of the education system as an institution that can be leveraged to experience achievement and upward mobility.

Predictors for Achievement of Asian or Pacific Islander Students

To calculate regression-based estimates of mean math achievement for the seventh subgroup of interest, students identifying as Asian or Pacific Islander, state-level variables representing factors related to parental level of education (%BA), the economic circumstances of students' families (FER), English language proficiency of students (%EL), and the quality of schooling (SQI) are again used as predictor variables. In addition, a predictor variable representing *Asian* students is used. This last predictor variable is operationalized as the percent of grade 8 Asian or Pacific Islander students by state who identify as Asian, but not Pacific Islander (*%A*).

Figure 2.4: Regression model for computing state-level synthetic estimates of mean math achievement of Asian / Pacific Islander students

$$\widehat{\boldsymbol{Y}}_{\overline{\boldsymbol{API}}_{\boldsymbol{math}}} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{o}} + \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}(\%BA) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}(FER) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{3}}(\%EL) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{4}}(SQI) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{5}}(\%A)$$

Asian (%A)

To disaggregate achievement results by subgroup, the NAEP program combines Asian students with Native Hawaiian and Pacific Islander students. Hence the term, Asian Pacific Islander (API). Native Hawaiian and Pacific Islander refers to persons who identify as Native Hawaiian, Samoan, Guamanian or Chamorro, Fijian, Tongan, or Marshallese and encompasses the people within the United States jurisdictions of Melanesia, Micronesia and Polynesia. The predictor variable, *%A*, is used to model the relationship between the percent of Asians that make up Asian Pacific Islander grade 8 populations within states and the mean math achievement of

Asian Pacific Islander 8th graders in each state (i.e., the NAEP-reported estimate for API students by state in the test sample).

Rationale for %A Predictor. While Asian or Pacific Islander residents in the United States have the highest rate of adults with a bachelor's or more advanced degree, when compared to other racial and ethnic groups, wide variability in educational attainment exists within the API subgroup by ancestral origin. Residents of the U.S. who identify as Asian generally attain higher levels of education compared to those who identify as Pacific Islanders. For instance, in 2015 the American Community Survey published statistics indicating that roughly half of residents who identified as Korean, Chinese or Japanese held a bachelor's or more advanced degree. By contrast, around fifteen percent of residents identifying as Hawaiian, Samoan or Fijian held a bachelor's or more advanced degree.

In a sense, Ogbu's (2003) theory on the academic orientations of involuntary and voluntary minorities can also be extended to the phenomenon by which Asians out-achieve Pacific Islanders. U.S. residents who identify as Pacific Islander tend to be Americans whose families have inhabited regions of what is today the United States for thousands of years. While many U.S. residents who identify as Asian come from families who have lived in the U.S. for many generations, a much larger proportion of Asians residing in the US are immigrants or first- and second-generation American.

Other research (Nisbett, 2009) points to differences in cultural beliefs and orientations to make sense of gaps in achievement between Asians and other groups. Confucianism, which still influences cultural beliefs and attitudes in parts of East Asia, promoted the idea that intelligence is acquired through hard work and personal effort. Nisbett (2009) argues that to this day, Asians

believe that intellectual accomplishment is primarily a matter of work, while other racial groups are more likely to believe intellectual accomplishment is more a matter of innate ability.

Predictors for Achievement of American Indian and Alaskan Native Students

To calculate regression-synthetic estimates of mean math achievement for the eighth subgroup of interest, students identifying as American Indian or Alaskan Native (AIAN), state-level variables representing factors related to parental level of education (%BA) and the economic circumstances of students' families (FER) are used as predictor variables. Only two predictor variables are used to model variation in mean math achievement of AIAN students across states because there are relatively few states for which State NAEP reports estimates of mean math achievement for the AIAN subgroup, which play role of target values in this study and represent outcome variable values in the regression model. For the test sample (grade 8 math in 2015), NAEP reported the mean estimates of AIAN students in just thirteen states.

The availability of mean math achievement estimates for AIAN students across states would be even fewer if not for an extra sampling effort undertaken by NAEP through a project known as NIES (National Indian Education Study). Every four years, the NAEP program conducts the NIES, which involves oversampling schools with relatively high proportions of AIAN students in select states to obtain more reliable and accurate estimates of AIAN achievement. Still, since AIAN students represent a small proportion of students nationally (about 1 percent), the NIES study is only able to obtain sufficiently large samples to meet reporting requirements for a relatively small number of states.

Figure 2.5: Regression model for computing state-level synthetic estimates of mean math achievement of American Indian / Alaskan Native students

$$\widehat{\boldsymbol{Y}}_{\overline{\boldsymbol{AINA}}_{\boldsymbol{math}}} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{o}} + \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}(\%BA) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}(FER)$$

Predictors for Achievement of Students of Two or More Races

To calculate regression-synthetic estimates of mean achievement for the ninth subgroup of interest, students identifying as two or more races, four state-level predictor variables representing previously described factors are used. These factors include parental level of education (*%BA*), the economic circumstances of students' families (*FER*), race and ethnicity of students (*%BHAIAN*), and school quality (*SQI*).

Figure 2.6: Regression model for computing state-level synthetic estimates of mean math achievement of students identifying with two or more races

$$\widehat{\boldsymbol{Y}}_{\overline{\boldsymbol{TP}}_{\boldsymbol{math}}} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{o}} + \widehat{\boldsymbol{\beta}}_{\mathbf{1}}(\%\text{BA}) + \widehat{\boldsymbol{\beta}}_{\mathbf{2}}(\text{FER}) + \widehat{\boldsymbol{\beta}}_{\mathbf{3}}(\%\text{B-H-AIAN}) + \widehat{\boldsymbol{\beta}}_{\mathbf{4}}(\text{SQI})$$

Predictors for Achievement of English Learners

To calculate regression-synthetic estimates of mean math achievement for the tenth and final subgroup of interest, students identifying as English learners, state-level variables representing factors related to parental level of education (*%BA* variable), the economic circumstances of students' families (*FER* variable), and school quality (*SQI* variable) are used as predictor variables.

Figure 2.7: Regression model for computing state-level synthetic estimates of mean math achievement of English learner students

$$\widehat{\boldsymbol{Y}}_{\overline{\boldsymbol{EL}}_{\boldsymbol{math}}} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{o}} + \widehat{\boldsymbol{\beta}}_{\mathbf{1}}(\%BA) + \widehat{\boldsymbol{\beta}}_{\mathbf{2}}(FER) + \widehat{\boldsymbol{\beta}}_{\mathbf{3}}(SQI)$$

The Case for a Hybrid Approach: FLEX CS

The third and final approach used for the estimation of states' mean math achievement across subgroups combines features of these first two techniques (MICE and FH)—an approach referred to in this dissertation as Flexible Cross-Survey Analysis (FLEX CS). Cross-Survey Analysis (CSA) refers to the combined analysis of data from different surveys. Use of CSA is

meant to increase the accuracy of parameter estimates, since combining estimates from different surveys results in increasing effective sample size, which should lower bias and uncertainty in parameter estimation (Magadin de Kramer, 2016). CSA is adopted in this study because there are multiple surveys and sources of data that can concurrently be used to estimate mean subgroup achievement on State NAEP.

The Case for Flexibility

A distinctive and appealing feature of this FLEX CS approach is that the final estimates are not required to be formed from the same subestimates. For instance, the estimate of mean achievement for one state's Hispanic students can be obtained through the combination of MICE and WPE subestimates, meanwhile the FLEX CS estimate for a different state's Hispanic students can be obtained through the combination of FH and NNI subestimates.

The appeal of this approach is that it involves combining only data from different sources that can reasonably be expected to improve prediction. Although combining estimates computed from different methods and from different sources is generally a helpful technique for improving accuracy, not all estimates should be expected to improve accuracy to the same degree. FLEX CS permits the researcher to select only those variables from the original data file (in this case, the test sample) that are most helpful for predicting missing values in dependent variables, while at the same time allowing the researcher to use helpful administrative data external to the original data file for prediction. This approach expands on the flexibility offered by MICE, in which select data variables from the original data file are used to impute values on a variable-by-variable basis, while simultaneously borrowing useful predictive data from other surveys, which is a common feature of the SAE framework and Cross-Survey Analysis.

Chapter 3: Methodology

To appreciate the scale of the research problem at hand it is helpful to examine Table 3.1, which displays the extent to which estimates of mean achievement on the 2015 grade 8 mathematics assessment are reported for subgroups across states. In the table, rows represent states and columns represent subgroups, and the presence of a dot (•) within a cell indicates that the mean math achievement estimate of the corresponding state subgroup is reported by NAEP. Conversely, empty cells indicate that the mean math achievement estimates for the corresponding subgroup and state are unreported. As an instructive example, for Alabama (AL), the state from the first row of Table 3.1, mean math achievement estimates are unreported for four separate subgroups, students who identify as Asian or Pacific Islander (API), American Indian or Alaskan Native (AIAN), mixed-race (TP), and as an English learner (EL).

The problem of missing achievement estimates is most acute for the race and ethnicity subgroups, especially the American Indian or Alaskan Native (AIAN) group. Estimates for students from this subgroup are only reported in 13 of the 50 states. Next, reporting is most sparse for students who identify as two or more races, followed by Asian Pacific Islander. Across states, the problem is most acute for Utah, which has missing estimates for 7 of 18 subgroups, followed by Maine and Vermont, each of which are missing estimates for 6 of 18 subgroups.

This dissertation attempts to determine whether it is justifiable to use any of three separate and progressively more complex methodological approaches to fill out matrices of estimates representing mean achievement on State NAEP, such as the one depicted in Table 3.1. In addition, this study tries to answer whether one of the three techniques generally outperforms the others and whether relative performance varies by subgroup.

Table 3.1: Unreported mean achievement estimates for grade 8 math 2015, State NAEP

State	FRL		PARENTAL EDUCATION				RACE & ETHNICITY						ENGLISH PROF.		LEARNING DISABILITY		GENDER	
	E	I	NHS	HS	SBA	BA	W	B	H	API	AIAN	TP	EL	NEL	SWD	NSWD	M	F
AL	•	•	•	•	•	•	•	•	•					•	•	•	•	•
AK	•	•					•	•	•	•	•	•	•	•	•	•	•	•
AZ	•	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•
AR	•	•	•	•	•	•	•	•	•				•	•	•	•	•	•
CA	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•
CO	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•
CT	•	•	•	•	•	•	•	•	•	•			•	•	•	•	•	•
DE	•	•	•	•	•	•	•	•	•	•				•	•	•	•	•
FL	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•
GA	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•
HI	•	•	•	•	•	•	•		•	•		•	•	•	•	•	•	•
ID	•	•	•	•	•	•	•		•					•	•	•	•	•
IL	•	•	•	•	•	•	•	•	•	•			•	•	•	•	•	•
IN	•	•	•	•	•	•	•	•	•			•	•	•	•	•	•	•
IA	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•
KS	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•
KY	•	•	•	•	•	•	•	•	•	•		•		•	•	•	•	•
LA	•	•	•	•	•	•	•	•	•					•	•	•	•	•
ME	•	•	•	•	•	•	•							•	•	•	•	•
MD	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•
MA	•	•	•	•	•	•	•	•	•	•			•	•	•	•	•	•
MI	•	•	•	•	•	•	•	•	•	•			•	•	•	•	•	•
MN	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
MS	•	•	•	•	•	•	•	•	•					•	•	•	•	•
MO	•	•	•	•	•	•	•	•	•					•	•	•	•	•
MT	•	•	•	•	•	•	•		•		•	•		•	•	•	•	•
NE	•	•	•	•	•	•	•	•	•			•		•	•	•	•	•
NV	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•
NH	•	•	•	•	•	•	•		•	•				•	•	•	•	•
NJ	•	•	•	•	•	•	•	•	•	•				•	•	•	•	•
NM	•	•	•	•	•	•	•		•		•		•	•	•	•	•	•
NY	•	•	•	•	•	•	•	•	•	•			•	•	•	•	•	•
NC	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
ND	•	•	•	•	•	•	•	•	•		•			•	•	•	•	•
OH	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•
OK	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•	•
OR	•	•	•	•	•	•	•		•	•		•		•	•	•	•	•
PA	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•
RI	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•
SC	•	•	•	•	•	•	•	•	•				•	•	•	•	•	•
SD	•	•	•	•	•	•	•		•		•			•	•	•	•	•
TN	•	•	•	•	•	•	•	•	•					•	•	•	•	•
TX	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•
UT	•	•					•		•		•		•	•	•	•	•	•
VT	•	•	•	•	•	•	•							•	•	•	•	•
VA	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•
WA	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
WV	•	•	•	•	•	•	•	•						•	•	•	•	•
WI	•	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•
WY	•	•	•	•	•	•	•		•		•			•	•	•	•	•
Missing	0	0	2	2	2	2	0	11	3	20	37	26	19	0	0	0	0	0

FRL = National free or reduced lunch program (*E* = Eligible, *I* = Ineligible); *NHS* = Did not finish high school, *HS* = Graduated high school, *SBA* = Some education after high school, *BA* = Graduated college; *W* = White, *B* = Black, *H* = Hispanic, *API* = Asian/Pacific Islander; *AIAN* = American Indian/Alaskan Native, *TP* = Two or more races; *English prof.* = English proficiency (*EL* = English learner, *NEL* = Not an English learner); *SWD* = Student with learning disability (including those with 504 plans), *NSWD* = Not a student with learning disability; *M* = Male student, *F* = Female student; Missing = total number of students for which NAEP does not report achievement by subgroup.

Overall Research Design and Methods

Evaluating the extent to which a technique performs well relative to other techniques is based on weighted Mean Absolute Error (wMAE)—a weighted measure of the distances between NAEP-reported estimates of mean achievement and estimates produced by the three techniques. This measure aggregates distances between mean estimates reported by NAEP and mean estimates produced through the techniques for cells where the NAEP estimate is available, by subgroup of interest. In addition to wMAE, accuracy is evaluated through a measure of coverage*,* which is calculated as a proportion and represents the frequency with which technique-produced estimates of mean math achievement lie within target intervals associated with corresponding NAEP-reported estimates. Hence, the denominator used for computing this proportion is a number that represents cells with NAEP-reported estimates.

It should be noted that use of the term *accuracy* instead of *bias* to refer to prediction error in this study is deliberate because the difference between technique-based estimates and achievement values reported by NAEP does not represent a pure measure of bias. The distinction stems from the fact that NAEP-reported achievement values are themselves estimates, and do not represent true population values (i.e., the actual mean math achievement of states' subgroups), and calculating bias would require knowing the actual achievement means.

Weighted Mean Absolute Error (wMAE)

The wMAE measure is calculated for each technique by pooling weighted prediction errors over states. It is computed for each technique *across* subgroups, which provides an overall measure of accuracy, as well as *per* subgroup, which permits inference regarding whether the relative accuracy of the techniques vary by subgroup.

The overall measure of accuracy (i.e., wMAE *across* subgroups) is calculated as follows in this study,

$$\text{Overall } wMAE_t = \frac{\sum_{i=1}^{376}((|\widehat{NAEP}_i - \widehat{Technique}_i|) \div SE_{\widehat{NAEP}_i})}{376},$$

where wMAE of a technique *t* is equal to a sum of weighted absolute differences (i.e., prediction errors) between the estimates for state subgroups *i* generated by the technique under study ($\widehat{technique}_i$) and the estimates for state subgroups *i* as reported by NAEP ($\widehat{NAEP}_i$), divided by the number of estimates made available by NAEP for subgroups of interest across states (*376*). Each absolute difference ($|\widehat{NAEP}_i - \widehat{Technique}_i|$) is weighted by the reciprocal of the standard error (i.e., estimated precision) associated with corresponding NAEP-reported estimates ($SE_{\widehat{NAEP}_i}$) before being summed.[15]

This weighting step has the desired effect of diminishing the relative contribution to the wMAE measure of absolute differences when standard errors associated with the NAEP-reported estimates ($SE_{\widehat{NAEP}_i}$) are relatively large. This way NAEP estimates of achievement that are calculated with less precision have less influence than NAEP estimates calculated with greater precision on the evaluation of techniques with regard to their relative predictive accuracy.

For the test sample used in this study, NAEP reports estimates for state subgroups in 776 instances. However, to limit the scope of this study, comparisons are made between mean estimates of math achievement produced from the three techniques to estimates reported by NAEP for subgroups of particular interest—those for which the NAEP program is unable to report direct estimates of all 50 states. The subgroups of interest include the four parental level of

[15] Estimates of standard error of technique-produced predictions of mean math achievement do not directly factor into calculation of wMAE, though these standard error estimates do have a central role in estimation with the FLEX CS technique. More detail on this topic is provided in the section of this chapter describing the FLEX CS procedure.

education subgroups, five race and ethnicity subgroups, and one English proficiency subgroup. These subgroups comprise a total of 376 NAEP-reported estimates across states and, hence, 376 points for comparison (i.e., target values). Thus, the denominator from the formula for $wMAE_t$ across subgroups of interest is equal to 376 for each technique evaluated in this study.

In addition to an evaluation of accuracy through wMAE across subgroups of interest, evaluation of accuracy through wMAE is conducted *per* subgroup of interest for each technique. This second set of comparisons helps address the third research question—*how techniques vary in their ability to predict achievement by subgroup*. The formulas used per subgroup, as expressed following this paragraph, vary only by the upper limit of summation and denominator that are used—both of which are equal to the number of states for which NAEP reports estimates of mean math achievement for the corresponding subgroup. For instance, NAEP does not report on the mean math achievement of students whose parents did not finish high school (NHS) for 2 of the 50 states and, thus, the upper limit of summation and denominator used for this subgroup is forty-eight (i.e., $50 - 2$).

$$NHS\ wMAE_t = \frac{\sum_{i=1}^{48}((|\widehat{NAEP}_i - \widehat{Technique}_i|) \div SE_{\widehat{NAEP}_i})}{48}$$

$$HS\ wMAE_t = \frac{\sum_{i=1}^{48}((|\widehat{NAEP}_i - \widehat{Technique}_i|) \div SE_{\widehat{NAEP}_i})}{48}$$

$$SBA\ wMAE_t = \frac{\sum_{i=1}^{48}((|\widehat{NAEP}_i - \widehat{Technique}_i|) \div SE_{\widehat{NAEP}_i})}{48}$$

$$BA\ wMAE_t = \frac{\sum_{i=1}^{48}((|\widehat{NAEP}_i - \widehat{Technique}_i|) \div SE_{\widehat{NAEP}_i})}{48}$$

$$B\ wMAE_t = \frac{\sum_{i=1}^{39}((|\widehat{NAEP}_i - \widehat{Technique}_i|) \div SE_{\widehat{NAEP}_i})}{39}$$

$$H\ wMAE_t = \frac{\sum_{i=1}^{47}((|\widehat{NAEP}_i - \widehat{Technique}_i|) \div SE_{\widehat{NAEP}_i})}{47}$$

$$API\ wMAE_t = \frac{\sum_{i=1}^{30}((|\widehat{NAEP}_i - \widehat{Technique}_i|) \div SE_{\widehat{NAEP}_i})}{30}$$

$$AIAN\ wMAE_t = \frac{\sum_{i=1}^{13}((|\widehat{NAEP}_i - \widehat{Technique}_i|) \div SE_{\widehat{NAEP}_i})}{13}$$

$$TP\ wMAE_t = \frac{\sum_{i=1}^{24}((|\widehat{NAEP}_i - \widehat{Technique}_i|) \div SE_{\widehat{NAEP}_i})}{24}$$

$$EL\ wMAE_t = \frac{\sum_{i=1}^{31}((|\widehat{NAEP}_i - \widehat{Technique}_i|) \div SE_{\widehat{NAEP}_i})}{31}$$

Coverage

Similar to wMAE, coverage statistics are calculated *across* subgroups and *per* subgroup of interest. To calculate coverage across subgroups of interest, let C(x) be the number of instances in which the technique-produced predicted values, $\{\widehat{Technique}_1, \ldots \widehat{Technique}_{376}\}$, fall within target intervals associated with corresponding NAEP-reported estimates of mean math achievement. Then, the coverage statistic across subgroups equals,

$$\frac{C(x)}{376}$$

Since the number of available target-values vary across subgroups of interest, so do the denominators in the previously expressed coverage formula, ($\frac{C(x)}{n}$), for calculating coverage statistics *per* subgroup of interest. For the parental level of education subgroups (NHS, HS, SBA, & BA), the denominator is equal to 48. For the Black (B), Hispanic (H), Asian Pacific Islander (API), American Indian/Alaskan Native (AIAN), two or more races (TP), and English learner (EL) subgroups the denominators equal 39, 47, 30, 13, 24, and 31, respectively.

The target intervals are expressed as,

$$(\widehat{NAEP}_{ij} \pm 0.2 * \widehat{NAEP\ sd}_I),$$

where $\widehat{NAEP}_{ij}$ represents the NAEP-reported estimate of mean math achievement for subgroup *i* in state *j* and $\widehat{NAEP\ sd}_I$ represents the median standard deviation estimate for subgroup *I* (i.e.,

the median value of the NAEP-reported state-level standard deviations for subgroup *I* across states). Thus the technique-produced estimate of mean math achievement for subgroup *i* in state *j*, $\widehat{Technique}_{ij}$, falls within its corresponding target interval if the absolute mean standardized difference, *b*, between $\widehat{Technique}_{ij}$ and $\widehat{NAEP}_{ij}$, is less than 0.2, where standard deviation is defined by $\widehat{NAEP}\ sd_I$. The *b* statistic is expressed as,

$$b = \frac{|\widehat{Technique}_{ij} - \widehat{NAEP}_{ij}|}{\widehat{NAEP}\ sd_I},$$

the absolute difference between technique-based and NAEP-reported estimates of mean math achievement for subgroup *i* in state *j*, divided by the median of the NAEP-reported state-level standard deviations for subgroup *I*. The denominator ($\widehat{NAEP}\ sd_I$) used for calculating the *b* statistic takes one of ten possible values in this study, one for each subgroup of interest, as follows,

$$b = \frac{|\widehat{Technique}_{NHSj} - \widehat{NAEP}_{NHSj}|}{31.5},$$

$$b = \frac{|\widehat{Technique}_{HSj} - \widehat{NAEP}_{HSj}|}{32.6},$$

$$b = \frac{|\widehat{Technique}_{SBAj} - \widehat{NAEP}_{SBAj}|}{30.6},$$

$$b = \frac{|\widehat{Technique}_{BAj} - \widehat{NAEP}_{BAj}|}{34.4},$$

$$b = \frac{|\widehat{Technique}_{Bj} - \widehat{NAEP}_{Bj}|}{33.4},$$

$$b = \frac{|\widehat{Technique}_{Hj} - \widehat{NAEP}_{Hj}|}{34.0},$$

$$b = \frac{|\widehat{Technique}_{APIj} - \widehat{NAEP}_{APIj}|}{38.1},$$

$$b = \frac{|\widehat{Technique}_{AINAj} - \widehat{NAEP}_{AINAj}|}{35.4},$$

$$b = \frac{|\widehat{Technique}_{TPj} - \widehat{NAEP}_{TPj}|}{35.2},$$

$$b = \frac{|\widehat{Technique}_{ELj} - \widehat{NAEP}_{ELj}|}{33.3},$$

In these formulas the acronym used for a subgroup (e.g., NHS) replaces "I" from the general formula and the actual median value of NAEP-reported state-level standard deviations for subgroup I replaces "$\widehat{NAEP}\ sd_I$".[16]

Reporting Results of Predictive Accuracy (wMAE & Coverage)

The results from computing wMAE and coverage are presented in chapter 5 in a table similar to Table 3.2 (provided below as an example), which demonstrates the wMAE and coverage statistics by technique for each subgroup of interest and then aggregated for all subgroups of interest.

[16] Denominator values are calculated from data gathered through the NAEP Data Explorer (NDE) tool hosted on the National Center for Educational Statistics (NCES) website-- https://www.nationsreportcard.gov/ndecore/landing. The R code used for calculating these values (median NAEP-reported state-level standard deviations per subgroup of interest) is provided on the author's GitHub page.

***Table 3.2*:** Subgroup and aggregate measures of wMAE and coverage by technique (template).

	MICE	FH	FLEX CS
Did not finish high school (n = 48)			
Weighted Mean Absolute Error (wMAE)	##.##	##.##	##.##
Coverage	.##	.##	.##
Graduated high school (n = 48)			
Weighted Mean Absolute Error (wMAE)	##.##	##.##	##.##
Coverage	.##	.##	.##
Some education after high school (n = 48)			
Weighted Mean Absolute Error (wMAE)	##.##	##.##	##.##
Coverage	.##	.##	.##
Graduated college (n = 48)			
Weighted Mean Absolute Error (wMAE)	##.##	##.##	##.##
Coverage	.##	.##	.##
Black (n = 39)			
Weighted Mean Absolute Error (wMAE)	##.##	##.##	##.##
Coverage	.##	.##	.##
Hispanic (n = 47)			
Weighted Mean Absolute Error (wMAE)	##.##	##.##	##.##
Coverage	.##	.##	.##
Asian/Pacific Islander (n = 30)			
Weighted Mean Absolute Error (wMAE)	##.##	##.##	##.##
Coverage	.##	.##	.##
American Indian/Alaskan Native (n = 13)			
Weighted Mean Absolute Error (wMAE)	##.##	##.##	##.##
Coverage	.##	.##	.##
Two or more races (n = 24)			
Weighted Mean Absolute Error (wMAE)	##.##	##.##	##.##
Coverage	.##	.##	.##
English learner (n = 31)			
Weighted Mean Absolute Error (wMAE)	##.##	##.##	##.##
Coverage	.##	.##	.##
Total (n = 376)			
Weighted Mean Absolute Error (wMAE)	##.##	##.##	##.##
Coverage	.##	.##	.##

In addition to Table 3.2, the values that factor into the calculation of accuracy statistics (i.e., technique-produced estimates of mean math achievement) are provided in a series of tables in Appendix B, similar to Table 3.3, presented on the following page as an example.

***Table 3.3*:** Example table of estimates (here, for students whose parents did not finish high school) by state and technique, including NAEP-reported estimates.

Did not finish high school (NHS)	NAEP Reported	MICE	FH	FLEX CS
AL	254 (2.5)	###	###	###
AK	--	--	--	--
AZ	269 (2.3)	###	###	###
AR	266 (2.4)	###	###	###
CA	261 (2.0)	###	###	###
CO	266 (2.6)	###	###	###
CT	254 (4.2)	###	###	###
DE	264 (2.8)	###	###	###
FL	264 (2.5)	###	###	###
GA	269 (2.4)	###	###	###
HI	270 (4.4)	###	###	###
ID	262 (2.6)	###	###	###
IL	270 (3.1)	###	###	###
IN	268 (3.1)	###	###	###
IA	261 (3.5)	###	###	###
KS	268 (4.1)	###	###	###
KY	260 (2.6)	###	###	###
LA	258 (2.5)	###	###	###
ME	267 (4.4)	###	###	###
MD	265 (3.4)	###	###	###
MA	267 (4.6)	###	###	###
MI	261 (3.7)	###	###	###
MN	275 (3.3)	###	###	###
MS	258 (3.0)	###	###	###
MO	257 (2.9)	###	###	###
MT	272 (3.7)	###	###	###
NE	262 (2.7)	###	###	###
NV	263 (2.0)	###	###	###
NH	269 (4.4)	###	###	###
NJ	267 (4.5)	###	###	###
NM	261 (2.2)	###	###	###
NY	267 (3.0)	###	###	###
NC	264 (2.5)	###	###	###
ND	266 (3.4)	###	###	###
OH	259 (4.6)	###	###	###
OK	263 (3.0)	###	###	###
OR	268 (2.4)	###	###	###
PA	261 (3.4)	###	###	###
RI	268 (2.5)	###	###	###
SC	271 (3.6)	###	###	###
SD	265 (4.1)	###	###	###
TN	265 (3.1)	###	###	###
TX	272 (1.9)	###	###	###
UT	--	--	--	--
VT	266 (3.6)	###	###	###
VA	268 (3.2)	###	###	###
WA	266 (3.0)	###	###	###
WV	255 (2.9)	###	###	###
WI	263 (3.8)	###	###	###
WY	272 (2.8)	###	###	###

Note: This tables omits estimates for 2 states (AK, UT), for which NAEP estimates aren't published for this particular subgroup.

Table 3.3 displays estimates of mean achievement and standard error from the test sample for one of the parental level of education subgroups ("Did not finish high school"; NHS) as reported by NAEP, as well as the estimates for each technique under study. Similar tables are provided in Appendix B for each subgroup of interest in this study.

Criteria for Recommending a Technique

Coverage

The coverage statistic calculated in this study plays the important role of signaling whether a technique could be recommended for use in practice—for instance, by NAEP researchers. A technique passes muster for recommendation in this study if, *across* subgroups of interest, at least 95 percent of the technique's predicted estimates of mean math achievement fall within corresponding target intervals, and *per* subgroup of interest, at least 80 percent fall within corresponding target intervals.

These criteria are selected based on results of a simulation analysis conducted for this dissertation with example data from the *EdSurvey* package (Bailey et al., 2019) in R. The package includes an example NAEP dataset of 16,915 rows, each representing a fictitious student with demographic and achievement information. In broad terms, the steps undertaken to conduct the simulation analysis involved computing target estimates of mean achievement and standard deviations per available subgroup from a random sample of 2,500 students from the full set of 16,915, a sample typical in size to samples used for each state in actual NAEP testing (U.S. Department of Education, 2002).[17] Then, repeatedly (1,000 times total) sampling from the remaining students in the example dataset, while setting the number of students from each

[17] Students who identify with two or more races, the "TP" subgroup, are not included in the example dataset. Thus, the simulation analysis included 9 of the 10 subgroups of interest.

subgroup to be sampled equal to the number sampled from the original group of 2,500 students, and calculating the proportion of 1000 sample means per subgroup falling within their corresponding target intervals.[18]

As demonstrated in Table 3.4, the overall coverage rate (i.e., across subgroups) was 0.97. By subgroup, 6 of 9 were greater than 0.90, while 8 of 9 had rates greater than 0.8. On the other hand, the coverage rate for one subgroup, representing American Indian and Alaskan Native (AIAN) students was much lower. Just about a third of sampled means for this subgroup were bound by their respective target interval.

Based on these results, despite the anomalous rate associated with the AIAN subgroup, it is reasoned that an overall coverage rate equal to or greater than 0.95 and a rate of at least 0.80 per subgroup represent markers of a successful technique in its ability to predict mean math achievement. The samples used for the simulation analysis are drawn from the same "population," the example dataset from the *EdSurvey* package (Bailey et al., 2019), with size of subgroup samples similar to what can typically be expected in NAEP testing. The simulation results should thus theoretically offer insight into how often sound predictions of mean math achievement can be expected to come within 0.2 standard deviations of NAEP-reported estimates of mean math achievement (i.e., fall within associated target intervals).

[18] The initial sample of 2,500 randomly drawn students to establish target intervals included at least 62 students per subgroup of interest, except for the AIAN subgroup. To resolve this issue, random sampling among AIAN students only was undertaken, which resulted in a target interval based on 70 AIAN students (see Table 3.4). The set of R code used for the simulation analysis is provided on the author's GitHub page.

***Table 3.4*:** Coverage rate results and additional statistics from simulation analysis

Subgroup	Estimand (target) mean	Estimand SD	n	Coverage rate
NHS	263	31.8	192	0.89
HS	266	30.6	446	0.99
SBA	278	33.0	445	1.00
BA	289	35.6	1091	1.00
B	255	32.6	491	0.99
H	261	33.5	360	0.99
API	293	36.8	110	0.80
AIAN	273	32.2	70	0.33
EL	241	34.7	143	0.96
TOTAL				**0.97**

Note: "n" statistics represent the number of sampled students per subgroup for calculating both estimand and estimator statistics.

wMAE

The other measure of accuracy, wMAE, is not used as criteria for making determinations in absolute terms—that, yes, a technique should be recommended for use, or that, no, it should not. On the other hand, it helps determine which techniques perform best in *relative* terms. If multiple techniques meet the criteria to be considered suitable for practical use, wMAE statistics help determine which performs *best* in terms of ability to accurately predict mean subgroup achievement.

The Three Techniques Used for Estimation of Mean Math Achievement

The three techniques used in this study are Multivariate Imputation by Chained Equations (MICE), the Fay-Herriot model (FH), and a Flexible Cross-Survey model (FLEX CS). The first approach (MICE) is form of Multiple Imputation, the second (FH) is a form of Small Area Estimation, and the third is a form of Cross-Survey analysis that combines features of MICE and FH, and provides the researcher flexibility in choice of data and model specification.

The approaches are progressively more complex in terms of the data that they require for prediction and the manner in which the predicted values are constructed. The MICE approach requires only test sample data, the NAEP-reported state-level estimates of mean math

achievement of 8th grade students in 2015. The FH technique requires restricted-use math achievement data of 8th grade students in 2015, as well as state-level administrative data. The FLEX CS technique draws on the data used for MICE and FH, as well as an additional set of district-level achievement data. The FH procedure involves combining two subestimates and FLEX CS involves combining up to four subestimates. It is presumed that additional layers of complexity enhance predictive performance.

Prediction with the MICE Technique

The MICE procedure is implemented with the *mice* package in R (van Buuren & Groothuis-Oudshoorn, 2011) and publicly available NCES data representing mean achievement of subgroups within states on grade 8 math in 2015 (i.e., the test sample).[19] Multiple Imputation, including *mice*, is traditionally used to impute missing values of incomplete variables of interest from a dataset, a total *m* times, which results in *m* complete datasets. For this study, however, the imputation procedure is separately administered for each target value within subgroups of interest. Hence, the MICE procedure is executed a total of 376 times, resulting in $376 \times m$ complete datasets.

This adaptation serves the evaluative nature of this study. The evaluation of predictive accuracy of techniques is based on comparisons between predicted and observed (i.e., target) values. By contrast, the technique is not applied to impute values into cells with actual missing data in this study. For each set of *m* imputations generated through MICE, one of the target values is withheld from the dataset *as if* it were missing. The MICE procedure is then executed and the average of *m* imputed values for the withheld target value is treated as the predicted

[19] The uppercase acronym notation of Multivariate Imputation by Chained Equations ("MICE") is generally used to reference or describe the technique as implemented in this study (i.e., the adaptation of the technique). On the other hand, the lowercase and italicized acronym ("*mice*"), is used to describe the technique in general, including how the technique is implemented in practice.

mean math achievement of the corresponding state's subgroup. This process repeats itself for each target value and, each time, a new target (observed) value of interest is withheld and the previously removed target value is returned to the dataset. To illustrate, Table 3.5 depicts this withholding process for the first and second administrations of the MICE technique in this study.[20] Cells that are color-coded dark gray demonstrate the location of withheld values.

Table 3.5*:* Depiction of "withholding" process (here, for the first and second administrations of MICE)

First 5 cases (sorted alphabetically) from test sample, *no target values withheld*

State	FRL		PARENTAL EDUCATION				RACE & ETHNICITY						ENGLISH PROF.		LEARNING DISABILITY		GENDER	
	E	I	NHS	HS	SBA	BA	W	B	H	API	AIAN	TP	EL	NEL	SWD	NSWD	M	F
AL	•	•	•	•	•	•	•	•	•					•	•	•	•	•
AK	•	•					•	•	•	•	•	•	•	•	•	•	•	•
AZ	•	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•
AR	•	•	•	•	•	•	•	•	•				•	•	•	•	•	•
CA	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

First 5 cases from test sample for *1st administration* of MICE. The NAEP-reported mean math achievement estimate of students from Alabama (AL) whose parents did not complete high school (NHS) is withheld.

State	FRL		PARENTAL EDUCATION				RACE & ETHNICITY						ENGLISH PROF.		LEARNING DISABILITY		GENDER	
	E	I	NHS	HS	SBA	BA	W	B	H	API	AIAN	TP	EL	NEL	SWD	NSWD	M	F
AL	•	•		•	•	•	•	•	•					•	•	•	•	•
AK	•	•					•	•	•	•	•	•	•	•	•	•	•	•
AZ	•	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•
AR	•	•	•	•	•	•	•	•	•				•	•	•	•	•	•
CA	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

First 5 cases from test sample for *2nd administration* of MICE. The NAEP-reported mean math achievement estimate of students from Arizona (AZ) whose parents did not complete high school (NHS) is withheld and the estimate of students from Alabama whose parents did not complete high school is returned.

State	FRL		PARENTAL EDUCATION				RACE & ETHNICITY						ENGLISH PROF.		LEARNING DISABILITY		GENDER	
	E	I	NHS	HS	SBA	BA	W	B	H	API	AIAN	TP	EL	NEL	SWD	NSWD	M	F
AL	•	•	•	•	•	•	•	•	•					•	•	•	•	•
AK	•	•					•	•	•	•	•	•	•	•	•	•	•	•
AZ	•	•		•	•	•	•	•	•	•	•		•	•	•	•	•	•
AR	•	•	•	•	•	•	•	•	•				•	•	•	•	•	•
CA	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

Note: For brevity, this table demonstrates just the first 5 of 50 cases from the test sample.

[20] This design is similar to the leave-one-out scheme commonly used in cross-validation research (LOOCV), though there are important differences. For instance, in LOOCV, "training" cases used for estimating regression coefficients may be discarded during linear regression if they have missing values. The mice algorithm always involves assigning temporary values to cases used for prediction where values are missing (see detail on this point in "Step 2"). In addition, LOOCV involves successively withholding each case from a dataset one time. By contrast, the MICE procedure in this study involves withholding NAEP-reported estimate values for variables (subgroups) of interest in the test sample.

Implementation of MICE also departs from its traditional use in that the values of interest in this study are the missing values themselves. By contrast, values of interest calculated with the technique are typically sets of pooled parameter estimates (e.g., regression coefficients) that are generated from sets of *m* complete datasets. Consider Figure 3.1, an edited version of a visual provided by van Buuren & Groothuis-Oudshoorn (2011), which communicates the main steps in multiple imputation with the *mice* package.

Figure 3.1: Stages in *mice*, emphasis on "imputed data" stage

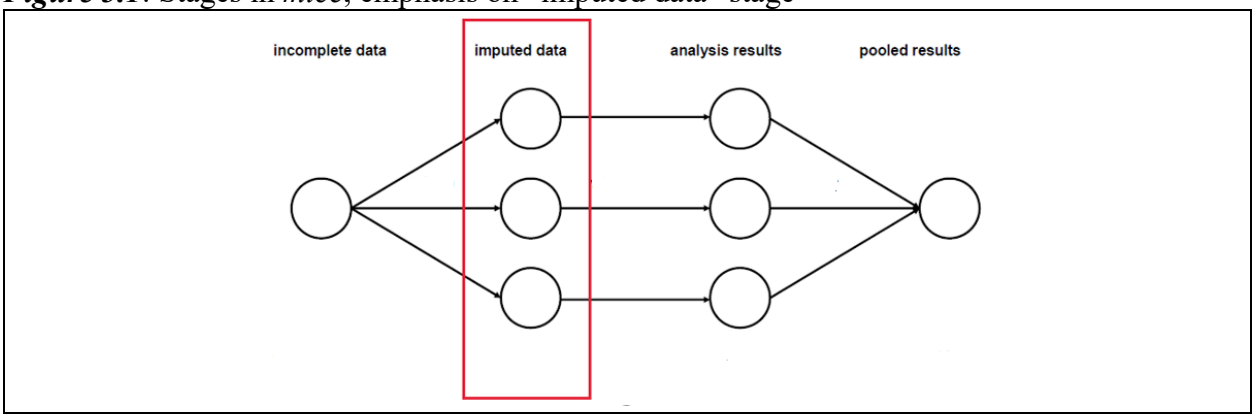


Figure 3.1 depicts a scenario in which a researcher starts from an incomplete dataset and specifies three sets of imputations ($m = 3$), which results in three separate complete data sets ("imputed data" stage). Next, the researcher analyzes each complete dataset separately and records parameter estimates of interest ("analysis results" stage). Finally, the estimates of interest are pooled in a manner that accounts for both the within- and between-imputation variance (van Buuren & Groothuis-Oudshoorn, 2011).[21] Attention is drawn to the "imputed data" stage in Figure 3.1 since the values of interest in this study are the imputed values themselves, which are

[21] Incorporating within- and between-imputation variance into the pooled variance estimate follows Rubin's rules (1987). This process entails *summing* three sources of variance—the between, the within and an additional source of sampling variance. The between represents the variance of a set of parameter estimates *across* imputed datasets. The within represents the arithmetic mean of sampling variance from each imputed dataset. The additional source of sampling variance is computed by dividing the between-variance estimate by the number of imputed datasets.

generated in the "imputed data" step. The adaptation of the technique in this study does not involve the subsequent stages ("analysis results" and "pooled results") depicted in Figure 3.1.

Step-by-step Procedure for Calculating Mean Achievement Estimates of Subgroups across States with MICE in this Study (the "Imputed Data" Step)

A total seven steps are outlined in subsequent paragraphs to describe how predicted values of mean subgroup achievement across states are computed with *mice* (van Buuren & Groothuis-Oudshoorn, 2011). To describe how the MICE procedure is implemented in this study, paragraphs in bulleted format are inserted following general descriptions of the *mice* procedure. Before delving into the steps, notation, definitions and context are documented.

***Table 3.6*:** Notation used to describe the MICE procedure

$\mathbf{X}$	A dataset with *n* cases (rows) and $\boldsymbol{p}$ variables (columns). • In this study, **X** corresponds to the test sample, a dataset with 50 rows (states) and 18 columns (subgroups), where cell values represent mean math achievement estimates for subgroups across states.
$\vec{\boldsymbol{p}}$	A vector of variables $\boldsymbol{p}$ from dataset **X** used for prediction. • In this study, $\vec{\boldsymbol{p}}$ represents the various sets of predictor variables from the chained regression equations. There are 10 such sets of predictors since there are 10 outcome variables to be imputed and hence 10 chained regression equations (more detail on these predictors is provided in Figure 3.2.
$\boldsymbol{p}_{inc}$	An incomplete variable from dataset **X**, which undergoes imputation. • In this study, $\boldsymbol{p}_{inc}$ represents any incomplete variable from the test sample. That is, any of the 10 subgroups of interest (the subgroups for which NAEP does not report an estimate of mean math achievement).
$\boldsymbol{p}$	Variables (columns) in dataset **X (**$\boldsymbol{p}$ can be either complete or incomplete). • In this study, there are 18 $\boldsymbol{p}$ variables—10 are incomplete and 8 are complete (not missing data).
$\boldsymbol{p_{inc}}^{\boldsymbol{obs}}$	Observed values from an incomplete variable from dataset **X** undergoing imputation. These are the values from each $\boldsymbol{p}_{inc}$ that are regressed on corresponding $\vec{\boldsymbol{p}}$**.**
$\boldsymbol{p}_{inc}{}_{\boldsymbol{-1}}^{\boldsymbol{obs}}$	An incomplete variable from dataset **X**, which undergoes imputation, with one target value withheld. • The withheld value serves as a target value to which the average of imputed values (for the corresponding cell) are compared. There are 376 such target values in the test sample (dataset **X**).
$\boldsymbol{y}_{\boldsymbol{p}_{inc}}$	Imputed (predicted) values for $\boldsymbol{p}_{inc}$.
β_o	The y-intercept term from a regression model.
$\vec{\boldsymbol{\beta}}$	A vector of regression coefficients. • In this study, there are 10 vectors of varying length (varying number of regression coefficients). For instance, the length of $\vec{\boldsymbol{\beta}}$ for imputing achievement values for the API subgroup is 2 (2 predictor variables), while the length of $\vec{\boldsymbol{\beta}}$ for imputing the BA subgroup is 9 (see Figure 3.2 for more detail).
$\overrightarrow{\boldsymbol{\beta} * \boldsymbol{p}}$	A vector of regression coefficients multiplied by associated variables $\boldsymbol{p}$ from dataset **X** used for prediction.
ε	Residual (error) term from a regression model.
σ^2	Variance of residuals from a regression model.
n	Cases (rows) in dataset **X**.
$\dot{}$	"Dot" placed above regression estimate or imputed value to indicate it is randomly sampled from a probability distribution.
t	Number of iterations (times the *mice* algorithm cycles across chained equations performing regressions and sampling estimates).
m	Number of complete datasets generated from the *mice* procedure. Therefore also the number of imputed values for each cell with missing data in **X** • In this study, *m* complete datasets are generated for 376 separate target values and the average of the imputed values for each (withheld) target value is the predicted value to which the target value is compared.

Context

Each incomplete variable $\boldsymbol{p}_{inc}$ from a dataset **X** is to be regressed on select predictor variables $\vec{\boldsymbol{p}}$ from dataset **X**. Variables $\vec{\boldsymbol{p}}$ used to predict each $\boldsymbol{p}_{inc}$ can be both complete or incomplete themselves. Each $\boldsymbol{p}_{inc}$ is separately specified to be regressed on select $\vec{\boldsymbol{p}}$ from **X**. This permits the regression method (e.g., logistic, linear) and predictor variables $\vec{\boldsymbol{p}}$ used to impute each $\boldsymbol{p}_{inc}$ to differ from one another.

- In this study, each $\boldsymbol{p}_{inc}$ is a continuous variable and imputations are created under the normal linear regression model, of the general form:

$$y_{p_{inc}} = \beta_o + \overrightarrow{\boldsymbol{\beta} * \boldsymbol{p}} + \varepsilon, \text{ where } \varepsilon \sim N(0, \sigma^2),$$

  where $\boldsymbol{y}$ represents predicted values for $\boldsymbol{p}_{inc}$ and $\overrightarrow{\boldsymbol{\beta} * \boldsymbol{p}}$ represents a vector of regression coefficients $\vec{\boldsymbol{\beta}}$ multiplied by values from variables $\vec{\boldsymbol{p}}$. Following guidance from Graham (2009), only variables $\boldsymbol{p}$ from **X** with moderate to high correlations with each $\boldsymbol{p}_{inc}$ are used as its predictors, $\vec{\boldsymbol{p}}$. For this study, moderate to high correlations are considered those greater than 0.5.[22] The dataset **X** (i.e., the test sample) is a 50-row (n) by 18-column ($\boldsymbol{p}$) data matrix, where rows n represent states and columns (i.e., variables) $\boldsymbol{p}$ represent demographic subgroups. Values in **X** represent NAEP-reported estimates of mean math achievement of 8th grade students in 2015. The test sample includes 18 $\boldsymbol{p}$ total—10 $\boldsymbol{p}_{inc}$ and 8 variables without missing data.

[22] The r > 0.5 criterion represents a rather arbitrary cut-point. Generally, Pearson product-moment correlations of 0.5 are considered moderate in strength. A more rigorous (i.e., higher) cut-point is used for the MICE procedure in the FLEX CS approach.

Specifying the Imputation Models

Step 1: Before the *mice* algorithm is executed, the order in which each $\boldsymbol{p}_{inc}$ is regressed on select $\vec{\boldsymbol{p}}$ is specified. This order is known as the "visiting sequence" (van Buuren, 2018). Although specifying a visiting sequence is not a requirement for running *mice* in R, a user-defined sequence can be helpful. Specifically, it can minimize the number of imputed values for each predictor $\boldsymbol{p}_{inc}$ that are randomly sampled from each $\boldsymbol{p}_{inc}{}'s$ corresponding observed values ($\boldsymbol{p}_{inc}{}^{obs}$) to begin the imputation procedure, which *mice* implements as default for *initializing* the algorithm.[23] In the event a visiting sequence is not specified prior to execution of the algorithm, the default in *mice* software (van Buuren & Groothuis-Oudshoorn, 2011) is to impute each $\boldsymbol{p}_{inc}$ in left-to-right order as they appear in dataset **X**.

- For this study, each $\boldsymbol{p}_{inc}$, a vector representing estimates of mean math achievement of subgroups across states, are regressed on at least two predictors in the order depicted in Table 3.7. Note that the outcome variables from the chain of regression equations displayed in Table 3.7 correspond to each $\boldsymbol{p}_{inc}$ from dataset **X**.

[23] More detail on *initializing* is provided in step 2.

Table 3.7: Visiting sequence of chained equations for this study

Order	Outcome variable	Predictor variable(s)
1	$*\overline{API}_{math}$	$\overline{SWD}_{math}, \overline{I}_{math}$
2	$*\overline{TP}_{math}$	$\overline{W}_{math}, \overline{NEL}_{math}$
3	$\overline{BA}_{math}$	$\overline{E}_{math}, \overline{I}_{math}, \overline{HS}_{math}, \overline{SBA}_{math}, \overline{W}_{math}, \overline{NEL}_{math}, \overline{NSWD}_{math}, \overline{M}_{math}, \overline{F}_{math}$
4	$\overline{SBA}_{math}$	$\overline{E}_{math}, \overline{I}_{math}, \overline{NHS}_{math}, \overline{HS}_{math}, \overline{BA}_{math}, \overline{NEL}_{math}, \overline{NSWD}_{math}, \overline{M}_{math}, \overline{F}_{math}$
5	$\overline{HS}_{math}$	$\overline{E}_{math}, \overline{I}_{math}, \overline{NHS}_{math}, \overline{SBA}_{math}, \overline{BA}_{math}, \overline{NEL}_{math}, \overline{NSWD}_{math}, \overline{M}_{math}, \overline{F}_{math}$
6	$\overline{B}_{math}$	$\overline{E}_{math}, \overline{NHS}_{math}, \overline{HS}_{math}, \overline{W}_{math}, \overline{H}_{math}$
7	$\overline{NHS}_{math}$	$\overline{E}_{math}, \overline{HS}_{math}, \overline{SBA}_{math}, \overline{BA}_{math}\ \overline{W}_{math}, \overline{B}_{math}, \overline{H}_{math}, \overline{AINA}_{math}, \overline{NEL}_{math}$
8	$\overline{H}_{math}$	$\overline{NHS}_{math}, \overline{B}_{math}, \overline{EL}_{math}$
9	$*\overline{AINA}_{math}$	$\overline{EL}_{math}, \overline{API}_{math}$
10	$\overline{EL}_{math}$	$\overline{H}_{math}, \overline{AINA}_{math}$

Note: The use of an asterisk (*) adjacent to the regression model for predicting values of the subgroup variables representing estimates of mean math achievement of students who identify as Asian or Pacific Islander (API), Two or more races (TP) and American Indian or Alaskan Native (AIAN) is meant to bring attention to the fact these response variables are regressed on at least one predictor variable with which they do not have a correlation of .50 or greater. These represent exceptions to the rule of "r > .5" and are specified to be regressed as such because these response variables do not have correlations with two other variables of at least 0.5. Instead, they are regressed on the two variables with which they have the highest correlations.

While there is no hard rule specifying a minimum number of cases per variable in linear regression (to avoid overfitting), it's common to ensure at least 8 to 10 cases be included per variable used for prediction (Tabachnik & Fidell, 2007; Altman, 1991). However, a more recent simulation study indicates that as little as two cases per predictor variable can be used in linear regression without inviting undue bias in parameter estimation (Austin & Steyerberg, 2015). For the MICE procedure used in this study, one predictor variable is permitted to enter a regression model for every 5 cases.

As an instructive example, let's consider the $\overline{BA}_{math}$ variable—the third p_{inc} to be imputed from the chained equations in Table 3.7. There are a total 47 cases used for regression for this outcome variable, so a maximum 9 predictor variables can be used. To expand on this example, consider that $\overline{BA}_{math}$ actually has a correlation of 0.5 or greater with 10 other variables

from the test sample (see Table 3.8). However, this outcome variable can only be regressed on 9 predictors and so the variable of 10 with which $\overline{\boldsymbol{BA}}_{\boldsymbol{math}}$ has the lowest correlation is removed. Thus, in this example, $\vec{\boldsymbol{p}}$ is a vector of 9 instead of 10 predictor variables.

Table 3.8: Pearson correlation matrix of NAEP-reported mean math scores from test sample

Subgroup	1	2	**3**	**4**	**5**	**6**	7	**8**	**9**	**10**	**11**	**12**	**13**	14	15	16	17	18
1. E																		
2. I	.68																	
3. NHS	.67	.47																
4. HS	.90	.75	.59															
5. SBA	.84	.79	.55	.83														
6. BA	.78	.90	.48	.86	.83													
7. W	.59	.87	.53	.70	.73	.88												
8. B	.54	.44	.57	.54	.41	.45	.55											
9. H	.38	.23	.59	.44	.27	.22	.27	.55										
10. API	.12	.45*	-.05	.11	.09	.39	.33	.22	-.08									
11. AIAN	-.28	.14	-.55	-.07	.04	-.05	-.01	-.89	.19	.22*								
12. TP	.23	.26	.38	.21	.19	.48	.65	.36	.34	.00	-.65							
13. EL	.20	.12	.42	.26	.06	.03	.03	-.14	.54	.20	.83	.33						
14. NEL	.81	.88	.54	.89	.87	.98	.87	.46	.24	.22	-.10	.49*	.08					
15. SWD	.68	.73	.39	.79	.66	.79	.67	.44	.32	.48*	.07	.13	.18	.78				
16. NSWD	.81	.87	.49	.88	.85	.95	.80	.41	.24	.36	-.04	.29	.11	.97	.82			
17. M	.82	.85	.49	.89	.84	.96	.80	.39	.24	.36	-.09	.39	.17	.97	.81	.98		
18. F	.81	.89	.53	.88	.88	.95	.83	.42	.27	.31	-.07	.30	.12	.97	.83	.98	.96	

Note: Bolded subgroup labels indicate subgroups of interest from the test sample (i.e., variables with missing data and for which regression models are constructed); Highlighted correlations indicate values equal or greater than .50; the asterisk (*) adjacent to ".48" in row 15 column 10 brings attention to the fact that the API subgroup variable does not have a correlation of .50 or greater with any other variable in the test sample. Instead it is regressed on the variable with which it has the highest correlation, which is the variable representing students with learning disabilities (r = .48). Asterisks are likewise placed next to the correlations between AIAN & API, and NEL & TP.

- In this study, once imputation models are specified, the test sample is manipulated just before executing the *mice* algorithm, a target value (i.e., NAEP-reported estimate of mean

math achievement) from the test sample is withheld. Thus, one of the ten separate $\boldsymbol{p}_{inc}$ from the chain of equations, henceforth $\boldsymbol{p}_{inc_{-1}}$, is missing an additional value.

Initializing the Mice Algorithm

Step 2: The *mice* algorithm involves the imputation of *initial* values into cells with missing data for each $\boldsymbol{p}_{inc}$ in **X** by randomly sampling with replacement from observed values from the corresponding $\boldsymbol{p}_{inc}$ ($\boldsymbol{p}_{inc}{}^{obs}$), which results in an initial complete dataset with no missing values.[24]

- In this study, for instance, consider the $\boldsymbol{p}_{inc}$ from test sample **X** representing the estimate of mean math achievement of students whose parents did not finish high school, $\overline{\boldsymbol{NHS}}_{\boldsymbol{math}}$. This $\boldsymbol{p}_{inc}$ has 2 missing values and 48 observed values, meaning the two missing values are imputed with values drawn at random and with replacement from the 48 observed values (when it is used as a predictor).
- Since values are drawn with replacement, the two imputed values for this $\boldsymbol{p}_{inc}$ can be the same, though this situation is improbable. However, for a $\boldsymbol{p}_{inc}$ with a greater proportion of missing data, randomly sampled observed values are likely, if not ensured, to repeat. Consider, for instance, the $\overline{\boldsymbol{AIAN}}_{\boldsymbol{math}}$ variable, which represents estimates of mean math achievement of students across states who identify as American Indian or Alaskan Native. This $\boldsymbol{p}_{inc}$ only has 13 observed values and hence 37 missing values, which means randomly drawn values from the set of 13 observed values must repeat during this initialization step of the *mice* procedure.

[24] Users of the *mice* package can also generate initial values through mean imputation, instead of randomly sampling (with replacement) observed values from the corresponding columns. The latter is the default approach in *mice*.

The Iterative Process

Step 3: After initialization, the vector of observed values (i.e., originally non-missing) from the first $\boldsymbol{p}_{inc}$ ($\boldsymbol{p}_{inc}{}^{obs}$) outlined in the visiting sequence is regressed on a pre-specified set of $\vec{\boldsymbol{p}}$. Hence, cases which originally had missing values for this first $\boldsymbol{p}_{inc}$ are not used for regression.

- In this study for instance, when target values from the API subgroup are withheld, 29 of 50 cases (*n*) from **X** are used for this regression, since NAEP did not report the mean math achievement estimate of 20 states for $\boldsymbol{Y}_{\overline{\boldsymbol{API}}_{\boldsymbol{math}}}$, the first $\boldsymbol{p}_{inc}$ outlined in the visiting sequence. Thus, in this study, given a target value from the API subgroup is withheld, the vector of observed values (i.e., originally non-missing) minus the case associated with the withheld value ($\boldsymbol{p}_{inc}{}_{\boldsymbol{-1}}^{\boldsymbol{obs}}$) is regressed on a pre-specified set of $\vec{\boldsymbol{p}}$**.**

Step 4: After this first regression model is fit and regression parameters are estimated, two important tasks are undertaken that ensure the *mice* procedure incorporates all sources of variability and uncertainty for each imputed value, a method in *mice* described by van Buuren (2018) as "prediction + noise + parameter uncertainty." This method takes two sources of uncertainty and variability into account for each imputed value. The first is uncertainty related to the estimated regression parameters—

$$\widehat{\beta_o}, \widehat{\overrightarrow{\boldsymbol{\beta}_{\boldsymbol{p}}}}, \text{ and } \widehat{\sigma^2},$$

where β_o represents the y-intercept, $\overrightarrow{\boldsymbol{\beta}_{\boldsymbol{p}}}$ represents a vector of regression coefficients associated with select $\vec{\boldsymbol{p}}$ from **X** on which the $\boldsymbol{p}_{inc}$ to be imputed is regressed, and σ^2 represents residual variance about the regression plane calculated from the regression fit. The second source is uncertainty about the regression planes, represented by the residual term "ε" from the regression formula—

$$\boldsymbol{y}_{\boldsymbol{p}_{inc}} = \beta_o + \overrightarrow{\boldsymbol{\beta} * \boldsymbol{p}} + \boxed{\varepsilon}, \text{ where } \boldsymbol{\varepsilon} \sim N(0, \sigma^2),$$

This error term represents the "noise" from the "prediction + noise + parameter uncertainty" moniker (van Buuren, 2018). Values ε represent the difference in observed and regression-generated (predicted) values, and are assumed to follow a normal distribution with a mean zero.

To account for the first source of uncertainty, related to regression coefficients ("parameter uncertainty"), *mice* uses a Bayesian framework for estimation with standard non-informative prior distributions. Following regression, the fitted coefficients for the $\boldsymbol{p}_{inc}$ criterion variable are replaced by random draws—$\dot{\beta}_o$, $\dot{\boldsymbol{\beta}}_{\boldsymbol{p}}$, $\dot{\sigma}^2$— from their respective estimated posterior distributions (van Buuren, 2018). The placement of dots above notation for these regression estimates signify that they are *randomly drawn* from their respective posterior distributions following regression fit. These sampled estimates ($\dot{\beta}_o$, $\dot{\boldsymbol{\beta}}_{\boldsymbol{p}}$, $\dot{\sigma}^2$) are used to generate conditional distributions for missing values in $\boldsymbol{p}_{inc}$. Establishing conditional distributions, which represent probability distributions about the regression planes and are assumed to approximate the normal distribution, permits the second source of uncertainty to be taken into account through random draws, $\dot{\boldsymbol{\varepsilon}}$, from these conditional distributions, which represent the values that are imputed. Hence, imputing values for $\boldsymbol{p}_{inc}$ involves:

1. $\boldsymbol{y}_{\boldsymbol{p_{inc}}^{obs}} = \beta_o + \overrightarrow{\boldsymbol{\beta} * \boldsymbol{p}} + \boldsymbol{\varepsilon}$, regressing $\boldsymbol{p_{inc}}^{\boldsymbol{obs}}$ on select $\vec{\boldsymbol{p}}$ variables.
2. $\dot{\beta}_o, \dot{\boldsymbol{\beta}}_{\boldsymbol{p}}, \dot{\sigma}^2 \sim \mathrm{P}(\widehat{\beta_o}, \widehat{\boldsymbol{\beta}_{\boldsymbol{p}}}, \hat{\sigma}^2 | \boldsymbol{p_{inc}}^{\boldsymbol{obs}}, \vec{\boldsymbol{p}})$, sampling regression estimates from the posterior distributions of estimates given regression fit of $\boldsymbol{p_{inc}}^{\boldsymbol{obs}}$ on select $\vec{\boldsymbol{p}}$ variables.
3. $\dot{\boldsymbol{\varepsilon}} \sim \mathrm{P}(\boldsymbol{\varepsilon} | \dot{\beta}_o, \dot{\boldsymbol{\beta}}_{\boldsymbol{p}}, \dot{\sigma}^2)$, sampling from a range of residual values about the regression plane conditioned on drawn regression estimates.

Hence the imputed value equals the value lying on the regression plane calculated from $\dot{\beta}_o$ and $\overrightarrow{\dot{\boldsymbol{\beta}}_{\boldsymbol{p}}}$, the predicted value, plus $\dot{\varepsilon}$, drawn from a range of values determined by the assumption of normality and $\dot{\sigma^2}$.

- In the context of this study, for instance, fitting the first regression model (i.e., imputing $\boldsymbol{Y}_{\overline{\boldsymbol{API}}_{\boldsymbol{math}}}$) results in a vector of two separate regression coefficient estimates ($\overrightarrow{\widehat{\boldsymbol{\beta}}_{\boldsymbol{p}}}$) assumed to follow a multivariate normal distribution, including their corresponding posterior distributions, since this linear regression model involves regressing $\boldsymbol{Y}_{\overline{\boldsymbol{API}}_{\boldsymbol{math}}}$ on two variables from test sample **X**. This means that the regression coefficients are randomly drawn from two separate posterior distributions, as well as random draws for the intercept and residual variance terms from their respective posterior distributions. The sampled estimates are then combined to estimate conditional distributions from which imputed values are drawn.

Step 5: After the 2-part imputation process (i.e., incorporating first "parameter uncertainty" and then "noise") is finished for the first $\boldsymbol{p}_{inc}$ from the chain of equations, as outlined in Figure 3.2, the *mice* algorithm moves onto the next equation and repeats the 2-part process. This time, however, if the previously imputed $\boldsymbol{p}_{inc}$ is used as one of the $\vec{\boldsymbol{p}}$ predictor variables, then the newly imputed values are used for performing regression instead of the values sampled during the initialization phase.

- In this study, for instance, the sampled values drawn through the 2-part process for the third $\boldsymbol{p}_{inc}$ ($\boldsymbol{Y}_{\overline{\boldsymbol{BA}}_{\boldsymbol{math}}}$) are used in the regression for the fourth $\boldsymbol{p}_{inc}$ from the visiting sequence ($\boldsymbol{Y}_{\overline{\boldsymbol{SBA}}_{\boldsymbol{math}}}$) since the set of $\vec{\boldsymbol{p}}$ on which $\boldsymbol{Y}_{\overline{\boldsymbol{SBA}}_{\boldsymbol{math}}}$ is specified to be regressed

includes $\boldsymbol{Y}_{\overline{\boldsymbol{BA}}_{\boldsymbol{math}}}$. For the remaining $\vec{\boldsymbol{p}}$ used for imputing $\boldsymbol{Y}_{\overline{\boldsymbol{SBA}}_{\boldsymbol{math}}}$, the values drawn during initialization are used.

Step 6: The 2-part imputation process subsequently continues for every other $\boldsymbol{p}_{inc}$ from the visiting sequence. The entire process of carrying out regressions across chained equations represents one of $\boldsymbol{t}$ iterations (or "cycles") of the *mice* algorithm. For any imputed value, the *mice* algorithm iterates (i.e., repeats the cycle) $\boldsymbol{t}$ times, and the values that are actually imputed for missing values in each $\boldsymbol{p}_{inc}$ from **X** represent those drawn from conditional distributions during the *final* iteration (the t^{th} iteration).

- Since there are 10 $\boldsymbol{p}_{inc}$ (including 1 $\boldsymbol{p}_{inc_{-1}}$ for each administration of MICE),[25] in **X** for this study and, hence, 10 chained equations, one iteration across the visiting sequence (following initialization) can be expressed as,

$$\dot{\beta_{o_1}}, \dot{\overrightarrow{\boldsymbol{\beta}_{\boldsymbol{p}_1}}}, \dot{\sigma^2}_1 \sim \mathrm{P}(\widehat{\beta_{o_1}}, \widehat{\overrightarrow{\boldsymbol{\beta}_{\boldsymbol{p}_1}}}, \widehat{\sigma^2}_1 \,|\boldsymbol{p}_{\boldsymbol{inc}_1^{obs}}, \vec{\boldsymbol{p}}_1)$$
$$\dot{y}_1^t \sim \mathrm{P}(y_1|\boldsymbol{p}_{\boldsymbol{inc}_1^{obs}}, \vec{\boldsymbol{p}}_1; \dot{\beta_{o_1}}, \dot{\overrightarrow{\boldsymbol{\beta}_{\boldsymbol{p}_1}}}, \dot{\sigma^2}_1)$$
$$\cdots$$
$$\cdots$$
$$\dot{\beta_{o_{10}}}, \dot{\overrightarrow{\boldsymbol{\beta}_{\boldsymbol{p}_{10}}}}, \dot{\sigma^2}_{10} \sim \mathrm{P}(\widehat{\beta_{o_{10}}}, \widehat{\overrightarrow{\boldsymbol{\beta}_{\boldsymbol{p}_{10}}}}, \widehat{\sigma^2}_{10} \,|\boldsymbol{p}_{\boldsymbol{inc}_{10}^{obs}}, \vec{\boldsymbol{p}}_{10})$$
$$\dot{y}_{10}^t \sim \mathrm{P}(y_{10}|\boldsymbol{p}_{\boldsymbol{inc}_{10}^{obs}}, \vec{p}_{10}; \dot{\beta_{o_{10}}}, \dot{\overrightarrow{\boldsymbol{\beta}_{\boldsymbol{p}_{10}}}}, \dot{\sigma^2}_{10})$$

For this study, $\dot{\beta_{o_1}}, \dot{\overrightarrow{\boldsymbol{\beta}_{\boldsymbol{p}_1}}}, \dot{\sigma^2}_1$ represent regression estimates drawn from their respective posterior distributions for imputing $\boldsymbol{Y}_{\overline{\boldsymbol{API}}_{\boldsymbol{math}}}$ (the first $\boldsymbol{p}_{inc}$ outlined in the visiting sequence), given observed values from $\boldsymbol{Y}_{\overline{\boldsymbol{API}}_{\boldsymbol{math}}}$ ($\boldsymbol{p}_{\boldsymbol{inc}_1^{obs}}$)—which play the role of

[25] In the first administration of MICE, as outlined in Table 3.5, the $\boldsymbol{p}_{inc_{-1}}$ is the subgroup representing students whose parents did not finish high school—the seventh of ten chained equations. Since one target value is removed from this subgroup of interest, the number of cases used for this regression is 47—one value less than the observed number of mean estimates of state achievement reported by NAEP for this subgroup. Note that for any 1 of 376 administrations of MICE, there are 10 chained equations (regression models), and that just one of the ten outcome variables being regressed is $\boldsymbol{p}_{\boldsymbol{inc}_{-1}^{obs}}$. The remaining nine are $\boldsymbol{p}_{inc}{}^{obs}$.

criterion values in regression and values from $\vec{\boldsymbol{p}}_{\mathbf{1}}$, values from a set of predictor variables on which $Y_{\overline{API}_{math}}$ is regressed. Similarly, $\dot{\beta_o}_{10}$, $\dot{\overrightarrow{\boldsymbol{\beta}_{\boldsymbol{p}_{\mathbf{10}}}}}$, $\dot{\sigma^2}_{10}$ are regression estimates drawn from their respective posterior distributions for imputing $\boldsymbol{Y}_{\overline{\boldsymbol{EL}}_{\boldsymbol{math}}}$ (i.e., the tenth and last $\boldsymbol{p}_{inc}$ outlined in the visiting sequence). It should be noted that, save for observed values in **X** (including ${\boldsymbol{p}_{inc}}^{obs}$), values for the remaining terms in the outlined cycle will vary across $\boldsymbol{t}$ iterations, as these values are successively re-estimated across iterations. For instance, coefficient values from $\overrightarrow{\boldsymbol{\beta}_{\boldsymbol{p}_{\mathbf{1}}}}$ will not be same at the second (t=2) iteration and third (t=3) iteration. Likewise values that are originally missing from **X** in $\vec{\boldsymbol{p}}_{\mathbf{1}}$ will differ across iterations. To continue with this example, values at t=3 change as a function of those drawn at t=2, as well as the stochastic process incorporated at t=3.

van Buuren and Groothuis-Oudshoorn (2011) indicate that 10 to 20 *t* iterations is sufficient to ensure imputations across *t* cycles converge around a similar value—meaning, values are imputed without too much bias. The default number of *t* iterations in *mice* software in R is set to 5 (van Buuren & Groothuis-Oudshoorn, 2011). After the first of *t* iterations, the two-part imputation process restarts,[26] once again beginning with the first regression from the visiting sequence. The sequence of values drawn across *t* iterations are referred to as "sampling streams" (van Buuren, 2018), hence the value that is actually imputed represents the value drawn at the end of the stream.

- For my study, *t* is set to 15. This is a relatively conservative specification, which prioritizes convergence over computational efficiency.

[26] Note that this "two-part procedure" refers to process by which two sources of uncertainty are taken into account (it does not refer to the seven separate steps used to communicate how the *mice* procedure works.

Note that the two-part procedure is executed during each of *t* iterations. This means 1) regression models are fit and parameters estimated, 2) regression parameters are sampled from the posterior distributions estimated from fitting the regressions, 3) the sampled parameters are used to generate conditional distributions, and 4) imputed values are drawn from conditional distributions.

Repeating Sets of Iterations

Step 7: Iterating through chained equations, as described in steps 3 to 6, results in one of *m* complete sets of data. Generally, at least 5 imputations per missing value are desired, meaning at least 5 sets of complete data are generated. Hence, steps 3 to 7 are typically repeated at least 5 times in practice. Fittingly, the default number of imputations used with *mice* software in R is set to 5 ($m = 5$). The imputations across *m* datasets vary because of the stochastic "two-part" procedure described in steps 4 through 6, where parameter estimates and predicted values are iteratively sampled at random from corresponding probability distributions.

Guidance around the appropriate *m* number of imputations to generate varies. Some research suggests *m* should be commensurate with the proportion of cases *n* with missing values in **X** (Von Hippel, 2009). Graham and colleagues (2007) demonstrate that estimation becomes more accurate the larger the *m*. Though the maximum *m* imputations evaluated by Graham and colleagues is 100 in their study, they suggest that choosing a number of *m* imputations is largely a matter of the computing power available for analysis.

- In this study *m* is set to 100, the largest specification of *m* evaluated by Graham and colleagues. While this number may appear excessive, the main reason for previously limiting *m* to a smaller number, limited computing power, is no longer a particularly compelling one. In addition, since the dataset (i.e., test sample) is relatively small (50 x

18 values), the number of *m* imputations can be conveniently increased beyond conventional sizes of *m* imputations without introducing too much computing burden.

Example Application of the MICE Procedure in this Study

As an instructive example, consider the first target value of interest from the test sample, the NAEP-reported estimate of mean math achievement of students in Alabama whose parents did not finish high school (AL/NHS). For this particular subgroup in Alabama, NAEP reports a mean math achievement estimate of 254. To predict this particular value of mean math achievement with MICE, this reported value (i.e., target value) of 254 is removed (only this value) prior to executing the *mice* algorithm, as previously depicted in Table 3.5.

The difference between the predicted value of achievement for this subgroup in Alabama, the average of *m* imputed values, and the NAEP-reported estimate of mean math achievement for this subgroup in Alabama (254) contributes to the calculation of wMAE. In more specific terms, the absolute difference divided by the standard error associated with the NAEP reported mean estimate of achievement (2.5) is summed together with similarly weighted absolute differences from other comparisons and divided by the number of comparisons.

To calculate coverage, this instance (i.e., comparison) of subgroup achievement on State NAEP (AL/NHS) is counted in the numerator of the coverage formula if the technique-produced (i.e., MICE-produced) value for AL/NHS falls within its corresponding target interval. The denominator represents the number of comparisons for which coverage is evaluated, which changes depending on whether coverage is examined across subgroups or for one subgroup in particular.

Verifying Credibility of MICE-produced Predicted Values

A common check following execution of the *mice* algorithm is to verify that the final imputations are credible. In general, an imputed value is credible if it could have been observed had it not been missing (van Buuren & Groothuis-Oudshoorn, 2011). In this study, to verify that MICE-based estimates of mean math achievement are credible, estimates of mean subgroup achievement predicted with MICE are compared to ranges of credible mean estimate values per subgroup. These ranges represent intervals of non-outlying values based on each subgroup of interest's distribution of NAEP-reported estimates of mean math achievement and Tukey's (1977) "$1.5 \times IQR$" rule for detecting outlying observations. Tukey's formula for calculating the lower and upper bounds of non-outlying observations are expressed as,

$$Q_1 - (1.5 \times IQR)$$
$$\&$$
$$Q_3 + (1.5 \times IQR),$$

where Q_1 and Q_3 represent the first and third quartiles (25th and 75th percentiles) of a set of observations and *IQR* (Interquartile Range) is the difference (i.e., distance) between the third and first quartiles (Q_3 and Q_1). Applying this formulation results in the reference ranges of credible mean achievement estimates presented in Table 3.9.[27]

Table 3.9: Lower- and upper-bounds of credible mean estimates per subgroup of interest

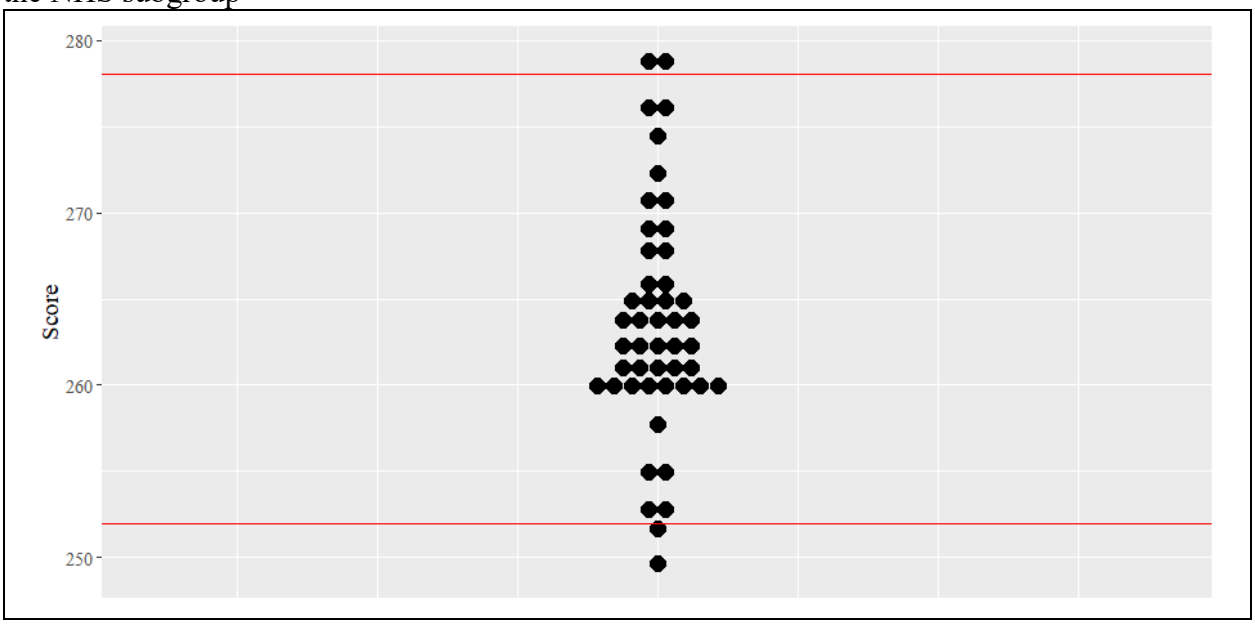
Subgroup of interest	Lower-bound	Upper-bound
NHS	252	278
HS	253	283
SBA	272	294
BA	276	310
B	245	274
H	257	282
API	276	335
AIAN	251	267
TP	267	297
EL	216	276

Note: Values are rounded to their nearest integer.

[27] The set of R Code used for computing ranges is documented on the author's GitHub page.

To aid in the identification of out-of-bound estimates of mean math achievement per subgroup of interest, the sets of predicted values produced from the MICE technique are graphed through dot plots with embedded lines that demarcate the lower and upper bounds of credible mean estimates for the respective subgroup. For the sake of illustration, consider the dot plot for the *NHS* subgroup from Figure 3.2. This plot displays the distribution of 48 values randomly drawn from a normal distribution with a mean of 265 and a standard deviation of 10. The red horizontal lines demarcate the lower and upper bounds of the proposed reference range for assessing the credibility of estimates of mean math achievement produced with MICE for the NHS subgroup. As can be observed, two values fall above the upper bound of 278 and two fall below the lower bound of 252.[28]

Figure 3.2: Dot plot of 48 *hypothetical* predicted values produced from the MICE technique for the NHS subgroup



NB: The red horizontal lines demarcate the lower and upper bounds, respectively 252 and 278, of the reference range used for assessing the plausibility of predicted values for the NHS subgroup.

[28] The set of R code, including seed, used for drawing values and generating the image from Figure 3.2 is provided on the author's GitHub page.

In the event actual MICE-based predicted values are out of range, such as the four out-of-bound simulated values depicted in Figure 3.2, then the MICE algorithm is modified. Specifically, the incomplete variables from the test sample to which the out-of-bound predicted values belong, are imputed through Predictive Mean Matching (PMM), while the remaining incomplete variables are still imputed through the normal linear regression model. Using PMM resolves the "out-of-bound" problem as this method restricts imputed values to draws from a set of observed values (Little, 1988). Implementation of PMM with the *mice* package in R involves forming a set of five candidate donors representing observed values from the corresponding column (variable) that are closest to the predicted value for the missing entry. Then, one of the five candidate donors is randomly drawn and used as the imputed value (van Buuren & Groothuis-Oudshoorn, 2011).

As an instructive example, imagine that one of the predicted values produced through MICE for a state from the AIAN subgroup is less than the lower bound of the proposed reference range (i.e., less than 251 for the AIAN subgroup). Then, the statistical code drafted for running the *mice* algorithm is amended so that the imputed values for the AIAN subgroup are based on PMM. This specification will no longer yield imputed values that fall below 251. Instead, this entry will be imputed by drawing, at random, from the five lowest observed values from the AIAN subgroup.

Prediction with the FH Technique

Estimating mean math achievement of subgroups across states with the Fay-Herriot (FH) model, a common technique from Small Area Estimation (SAE), involves borrowing information from area-level administrative data to improve direct estimates of area-level statistics of interest. To "improve" a direct estimate with a model-based estimate in the SAE framework generally

means to render an estimate more efficient. In this study, "area" corresponds to states. Hence, state-level administrative data are used to improve direct estimates of mean math achievement of demographic subgroups aggregated to the state level. Implementation of the FH model in this study can be expressed as follows,

$$\hat{\delta}_{ij}^{EBLUP} = \hat{\gamma}_{ij}\hat{\delta}_{ij}^{DIR} + (1-\hat{\gamma}_{ij})x_{ij}^{\mathrm{T}}\hat{\beta},$$

Where the mean math achievement estimate of subgroup *i* in state *j* is an Empirical Best Linear Unbiased Predictor ($\hat{\delta}_{ij}^{EBLUP}$), a precision-weighted combination of the direct estimate of mean math achievement of subgroup *i* in state *j* (i.e., the design-based estimate) and a regression estimator (i.e., a model-based estimate) of mean math achievement of subgroup *i* in state *j* ($x_{ij}^{\mathrm{T}}\hat{\beta}$), sometimes referred to as the "indirect" or "synthetic" estimate.

The use of the term BLUP, instead of BLUE, indicates that the statistical model used for approximating the parameter value of interest is a mixed-effects model (Galwey, 2014). By contrast, the term BLUE is more frequently used in statistics to describe estimates from fixed-effects models, such as the Ordinary Least Squares (OLS) regression model. The term *Empirical* relates to Empirical Bayes (EB) methods, a set of statistical procedures similar to full Bayesian methods, but where prior distributions are estimated from the observed (empirical) sample of data. Precision-weighted combinations of estimates, such as the EBLUPs produced with the FH technique, are characteristic of estimates computed with EB methods (Braun & Jones, 1984).

Computing Direct Estimates

The direct estimates ($\hat{\delta}_{ij}^{DIR}$) are computed from small random samples of students from subgroups of interest within states for which NAEP *is able to* report estimates of mean achievement. Put differently, the random samples are drawn from cases representing students identified with subgroups of interest in states for which NAEP *was able* to sample at least 62

students.[29] The random samples are drawn from restricted-use data from the National Center for Education Statistics, with sample size varying by subgroup. The size of the samples drawn is set to the median sample size of students available from the restricted-use data for the respective subgroup of interest in states that *do not* meet the rule-of-62. The steps involved in computing direct estimates are provided in finer detail in the next several paragraphs.

Step 1: Set target values equal to State NAEP-reported estimates of mean math achievement for subgroups of interest. These are the subgroups for which reporting is incomplete. In the test sample, there are 376 such target values (see table 3.1).

Step 2: For each of the 376 target values, each of which correspond to a different state subgroup, draw a random sample of students of size n from a subset of the restricted-use NAEP dataset for grade 8 math results in 2015. The subset is restricted to students used by NAEP for computing and reporting estimates of mean math achievement for the corresponding state subgroup. Since State NAEP results are based on public school students (i.e., non-charter and charter), this step involves removing private school students in addition to students that do not form part of the state and subgroup pair of interest.

The size of the random sample for each state subgroup, as described, varies by subgroup and is set to the median number of students sampled by NAEP from subgroups in states that are not reported by NAEP. This decision permits the simulation of scenarios in which researchers have relatively small samples of students (e.g., $n < 62$) from which to compute direct estimates of mean math achievement. Setting n to the median sample size of students from unreported subgroups is deliberate as this sample size represents a typical number of students that

[29] While it would be more purposeful *in practice* to use SAE to estimate the mean math achievement of subgroups that NAEP does not report, this approach is impractical for the focus of this study. The measures of accuracy on which the techniques examined in this dissertation are evaluated requires the use of target values to which predicted values can be compared.

researchers can expect to have available in practice when attempting to compute estimates for low-incidence populations (e.g., Black students in Vermont).[30]

As an instructive example, consider the race and ethnicity subgroup representing Hispanic students, for which the NAEP program does not report estimates of mean math achievement for three states (Maine, Vermont & West Virginia) in the test sample. The sample size used for computing the direct estimate ($\hat{\delta}_{ij}^{DIR}$) for this subgroup across states is equal to the median number of Hispanic students available across these three states. Hypothetically, if the number of Hispanic students in the test sample in Maine, Vermont, and West Virginia are 20, 40, and 10, then the sample size used for computing the direct estimate for this subgroup is 20.

Step 3: With each randomly drawn sample, each corresponding to one of the 376 state-subgroup pairs, compute a direct estimate of mean math achievement and standard error that accounts for NAEP's complex sampling design. This includes making appropriate use of sampling weights associated with each student, which are calculated based on students' sampling stratum (i.e., state) and cluster (i.e., school), as well as use of all plausible values drawn from estimated posterior distributions for each sampled student's grade 8 math ability.

Student-level NAEP data are nested. As such, analysis of these data require the researcher to model the dependency that exist among data from the same cluster. The failure to account for the nested nature of these data risks calculating standard errors of mean achievement that are downwardly biased (O'Dwyer & Parker, 2014). Appropriate use of plausible values involves calculating the statistic of interest (i.e., mean subgroup achievement in this study) separately for each set of plausible values and then pooling results. By contrast, it is

[30] This 'median sample size' specification presents a limitation to the inferences that can be made about the FH approach's general utility. In practice, researchers will have access to samples that do not meet the rule-of-62 that are at times smaller than the proposed median sample and at other times greater than the proposed median sample.

inappropriate to first average all plausible values for each individual student and then calculate the statistic of interest. Similar to the failure to account for clustered data, the latter approach to handling plausible values produces mean variance estimates that are unduly small (von Davier, Gonzalez & Mislevy, 2009).

Computing Regression-Synthetic (Model-Based) Estimates

The synthetic estimates of mean subgroup achievement across states ($x_{ij}^{\mathrm{T}}\hat{\beta}$) are computed with sets of 10 separate regression models, one per subgroup of interest, using Ordinary Least Squares (OLS) estimation. The number of regression models per subgroup set varies with the number of target values available per subgroup. For instance, there are 48 regression models used for the first subgroup of interest from Figure 3.3 (NHS), since there are 48 NAEP-reported estimates of mean math achievement for this subgroup.[31] In total, there are 376 regression models used for computing regression-synthetic estimates—one per state-and-subgroup pair of interest.

The reason for so many separate regression models is that, for any subgroup set of regressions, the values from the criterion variable are made up of the corresponding NAEP-reported estimates of mean achievement, as well as a unique direct estimate computed from one of the small ($n < 62$) samples randomly drawn from restricted-use data. The latter estimate replaces its corresponding NAEP-reported estimates in the criterion variable.

[31] Note that unlike the MICE procedure, regressions in the FH approach include the cases associated with the target value being predicted. Hence, for instance, 48 cases are used for regressing the NHS outcome variable on select predictors in the FH approach, but 47 cases are used for this outcome variable in the MICE procedure. To help make sense of this difference, consider that MICE deals with a missing data problem, whereas the FH technique involves improving a direct estimate that is calculated from a small sample. More detail on the specification and implementation of the regression models used in the FH approach is offered in "The EBLUP" section that follows, as well as Appendix A.

In other words, for each of the 376 regression models, the values from the criterion variable consist of one of the direct estimates computed from the small random samples and the remaining are NAEP-reported estimates of mean math achievement. For instance, in the first of 376 regressions, the small sample mean estimate computed through direct estimation, as outlined in the previous section, for the subgroup representing students whose parents did not finish high school ($\overline{\boldsymbol{NHS}}_{\boldsymbol{math}}$) in Alabama is the value from the outcome variable $\widehat{\boldsymbol{Y}}_{\overline{\boldsymbol{NHS}}_{\boldsymbol{math}}}$ for Alabama from Figure 3.3, and the remaining values for $\widehat{\boldsymbol{Y}}_{\overline{\boldsymbol{NHS}}_{\boldsymbol{math}}}$ in Figure 3.3 are those published by State NAEP. Together, these values are regressed on state-level variables representing percent of students identified as Black, Hispanic or American Indian/Alaskan Native (%B-H-AIAN), a measure of median family income and wealth (FER), percent of students identified as English learners (%EL), and a measure of school quality (SQI).

Figure 3.3: Regression equations for computing regression-synthetic (model-based) estimators

$$\widehat{\boldsymbol{Y}}_{\overline{\boldsymbol{NHS}}_{\boldsymbol{math}}} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{o}} + \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}(\%\text{B-H-AIAN}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}(\text{FER}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{3}}(\%\text{EL}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{4}}(\text{SQI})$$

$$\widehat{\boldsymbol{Y}}_{\overline{\boldsymbol{HS}}_{\boldsymbol{math}}} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{o}} + \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}(\%\text{B-H-AIAN}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}(\text{FER}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{3}}(\%\text{EL}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{4}}(\text{SQI})$$

$$\widehat{\boldsymbol{Y}}_{\boldsymbol{SBA}_{\boldsymbol{math}}} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{o}} + \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}(\%\text{B-H-AIAN}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}(\text{FER}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{3}}(\%\text{EL}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{4}}(\text{SQI})$$

$$\widehat{\boldsymbol{Y}}_{\overline{\boldsymbol{BA}}_{\boldsymbol{math}}} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{o}} + \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}(\%\text{B-H-AIAN}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}(\text{FER}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{3}}(\%\text{EL}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{4}}(\text{SQI})$$

$$\widehat{\boldsymbol{Y}}_{\overline{\boldsymbol{B}}_{\boldsymbol{math}}} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{o}} + \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}(\%\text{BA}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}(\text{FER}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{3}}(\%\text{AA}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{4}}(\text{SQI})$$

$$\widehat{\boldsymbol{Y}}_{\overline{\boldsymbol{H}}_{\boldsymbol{math}}} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{o}} + \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}(\%\text{BA}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}(\text{FER}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{3}}(\%\text{MX}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{4}}(\%\text{EL}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{5}}(\text{SQI})$$

$$\widehat{\boldsymbol{Y}}_{\boldsymbol{API}_{\boldsymbol{math}}} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{o}} + \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}(\%\text{BA}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}(\text{FER}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{3}}(\%\text{A}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{4}}(\%\text{EL}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{5}}(\text{SQI})$$

$$\widehat{\boldsymbol{Y}}_{\overline{\boldsymbol{AINA}}_{\boldsymbol{math}}} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{o}} + \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}(\%\text{BA}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}(\text{FER})$$

$$\widehat{\boldsymbol{Y}}_{\overline{\boldsymbol{2+}}_{\boldsymbol{math}}} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{o}} + \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}(\%\text{BA}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}(\text{FER}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{3}}(\%\text{B-H-AIAN}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{4}}(\text{SQI})$$

$$\widehat{\boldsymbol{Y}}_{\overline{\boldsymbol{EL}}_{\boldsymbol{math}}} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{o}} + \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}(\%\text{BA}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}(\text{FER}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{3}}(\text{SQI})$$

Covariate labels: *%B-H-AIAN* = state percent of students identified as Black, Hispanic or American Indian/Alaskan Native; *FER* = composite measure of states' median family income and wealth; *%EL* = state percent of students identified as English learners; *SQI* = measure of school quality in the state; *%BA* = state percent of adults with at least a bachelor's degree; *%AA* = state percent of Black population identified as African American; *%MX* = state percent of Hispanic population of Mexican descent; *%A* = state percent of grade 8 Asian and Pacific Islander students identified as Asian.

While it may appear more useful to set all values from the outcome variables equal to NAEP-reported estimates of mean math achievement, as opposed to sequentially replacing target values of interest with small sample direct estimates, this design does suit the evaluative nature of this study. The proposed design imitates a scenario in which NAEP researchers are confronted with the task of computing estimates of mean achievement with small samples ($n < 62$). In addition, the NAEP-reported values play the role of target values in this study and thus also using these NAEP-reported estimates as the direct estimates from the FH models would result in FH-produced estimates of mean math achievement that are unduly close to the target values. That is, setting all of the direct estimates of state-subgroup pairs of interest in the FH models to NAEP-reported estimates would unfairly favor the predictive performance of the FH technique, relative to the other predictive techniques under evaluation.

It could also appear more useful to set the values from criterion variables all equal to the direct estimates computed from the small samples drawn from restricted-use data. However, this sort of specification would fail to leverage the strength of the NAEP-reported estimates. The direct estimates from small random samples would be less accurate than the NAEP-reported estimates, which would result in more biased estimation of the relationships between criterion and predictor variables. Ultimately, this specification raises the chances of computing less accurate regression-synthetic estimates of mean math achievement.

In brief, the process by which each synthetic-regression estimate is calculated can be described in the following few steps:

1. A criterion variable representing NAEP-reported estimates of mean math achievement is set to be regressed on a group of predictor variables, as specified in Figure 3.3.

2. Prior to fitting the regression model, one of the criterion values is replaced with its corresponding estimate of mean math achievement computed from a small sample (i.e., n < 62) randomly drawn from restricted-use data.
3. The value predicted from the regression fit for the case (state) associated with the replacement estimate is the regression-synthetic estimate for the corresponding state.

These steps are repeated for each target value from the test sample, a total 376 times. Each time a different NAEP-reported estimate of mean math achievement from the criterion variable is replaced with its corresponding small sample direct estimate of mean achievement.

The EBLUP

The FH-produced estimate of mean math achievement (the EBLUP) is a precision-weighted combination of the direct and synthetic regression estimates. As an instructive example, consider a state whose direct estimate of mean math achievement ($\hat{\delta}_{ij}^{DIR}$) equals 260.0 with variance of 10.0 and whose synthetic estimate of mean math achievement ($x_{ij}^{\mathrm{T}}\hat{\beta}$) equals 265.0 with variance of 5.0. The former variance is obtained by calculating the variance estimate of the sample used for direct estimation. The latter variance is equal to the mean squared error (MSE) statistic computed from fitting the regression, which reflects the variance of residuals (error terms about the regression plane). In this example, total variance is equal to 15.0 and thus the proportion of total variance attributable to the regression estimator ($\hat{\gamma}_{ij}$) is one-third (i.e., 5.0/15.0) and the proportion attributable to the direct estimator (1-$\hat{\gamma}_{ij}$) is two-thirds (10.0/15.0). Plugging these numbers into the right side of Formula A results in Equation A and an EBLUP of about 263.3.

Formula A

$$\hat{\delta}_{ij}^{EBLUP} = \hat{\gamma}_{ij}\hat{\delta}_{ij}^{DIR} + (1-\hat{\gamma}_{ij})x_{ij}^{\mathrm{T}}\hat{\beta}$$

Equation A

$$\hat{\delta}_{ij}^{EBLUP} = 1/3(260.0) + 2/3(265.0)$$

Intuitively this result ($\hat{\delta}_{ij}^{EBLUP}$= 263.3) makes sense, given the regression estimate is calculated with greater precision compared to the direct estimate. The variance of the regression estimate (5.0) is smaller than the variance of the direct estimate (10.0). As a result, the EBLUP (263.3) comes closer to the regression estimate of mean math achievement (265.0) than the direct estimate of mean math achievement (260.0).

Calculation of the EBLUPs is implemented with the *sae* package in R (Molina & Marhuenda, 2015), which provides a variety of functions for Small Area Estimation, including the FH method. The calculation of direct estimates from restricted-use student-level data is implemented through Stata v16.1 and the *svy* package (2019).[32] The *svy* package includes functionality to analyze complex survey data, such as achievement data from State NAEP, and permits users to incorporate each student's sampling weight (*ORIGWT*), jackknife replicate weights for their cluster (*SRWT'*s), and plausible values of grade 8 math achievement (*MRPCM*'s).[33] The direct estimates computed with Stata are incorporated in the calculation of EBLUPs with the *sae* package in the manner described in the previous section on calculating regression-synthetic estimates, whereby direct estimates from small random samples, computed

[32] Using the *EdSurvey* package (Bailey et al., 2019) in R was first proposed to compute direct estimates. The *EdSurvey* package includes functionality to analyze complex survey data and was intentionally designed for the analysis of education data from the National Center for Education Statistics (NCES), including data from State NAEP. Complications related to COVID-19 compelled the use of a software environment incompatible with the *EdSurvey* package.

[33] The capitalized and italicized text in parentheses reflect the naming of the variables as they appear in the NCES restricted-use data set. Example code used for computing the direct estimates is provided in Appendix A. The full set of code used for computing direct estimates is provided on the author's GitHub page.

with restricted-use data, sequentially replace values from criterion variables from regression models.

Prediction with the FLEX CS Technique

The third and final technique used for estimating mean subgroup achievement, Flexible Cross-Survey Analysis(FLEX CS), draws on features of the first (MICE) and second (FH) techniques. Similar to the MICE technique, FLEX CS uses *in-sample* data as predictor variables to support estimation of values of interest. Similar to the FH approach, FLEX CS combines estimates (*subestimates*) from different sources of data to calculate final estimates. FLEX CS is described as a cross-survey approach for the technique's emphasis on combining data from different sources for parameter estimation.

The subestimates that form the FLEX CS estimates are computed from four different techniques—1) MICE , 2) FH, 3) a Weighted Poststratified Estimator (WPE) calculated with district-level estimates of achievement from the Stanford Education Data Archive (Reardon et al., 2017), and 4) Nearest-Neighbor Imputation (NNI).

A distinct feature of this FLEX CS approach is that the final estimates are not required to be formed from the same subestimates. For instance, the estimate of mean math achievement for one state's Hispanic students may be computed as the combination of FH and WPE subestimates, meanwhile the FLEX CS estimate for a different state's Hispanic students may be computed as the combination of FH and NNI subestimates.

This flexibility is built into the approach to permit only the combining of subestimates that are justifiably presumed to be accurate estimators of a particular state's subgroup. In other words, subestimates from the four separate techniques (MICE, FH, WPE, NNI) are only used or combined if there exists evidence to suggest that the information used in the approach could

support accurate prediction. As an instructive example, NNI might be used to estimate the achievement of a state's subgroup if an estimate for the same subgroup is available in a very similar state (e.g., South Dakota & North Dakota). On the other hand, if a state's nearest neighbor is not particularly similar, then NNI would not be used to predict mean math achievement for a subgroup within that state. The prevailing principle that guides model specification in FLEX CS estimation is that predictor variables that are presumably unhelpful for predicting values of variables of interest should not influence estimates of mean achievement.

While the flexibility in selection of subestimates to be combined is presumed to support accurate estimation of values of interest, it should be noted that a drawback to this flexibility is an inability to express FLEX CS as a standard model. Put differently, the FLEX CS technique cannot be expressed as the combination of a *specific* set of estimates. This aspect of the approach makes it challenging, for instance, to apply FLEX CS as presented in this study to other research problems.

Criteria for Using a Subestimate in the FLEX CS Technique

1. **MICE Subestimate**.

The MICE procedure is used for estimating mean math achievement of subgroups of interest if at least two auxiliary variables, which serve as predictor variables in the MICE equations, have a correlation with the response variable of at least .80. Research suggests that using auxiliary variables more highly correlated with variables to be imputed is generally associated with greater reduction in bias (Graham, 2009; Johnson & Young, 2011) and a correlation of .80 or higher is considered to represent a strong relationship between variables (Taylor, 1990). Using this criterion results in the removal of several predictor variables from the specification of MICE described earlier, as demonstrated across Tables 3.10 and 3.11.

Table 3.10: Visiting sequence from first implementation of MICE procedure with predictor variables struckthrough that do not have a correlation of at least 0.80 with the response variable

Order	Outcome variable	Predictor variable(s)
1	$*\overline{API}_{math}$	~~$\overline{SWD}_{math}$, $\bar{I}_{math}$,~~
2	$*\overline{2+}_{math}$	~~$\overline{W}_{math}$, $\overline{NEL}_{math}$~~
3	$\overline{BA}_{math}$	~~$\bar{E}_{math}$~~, $\bar{I}_{math}$, $\overline{HS}_{math}$, $\overline{SBA}_{math}$, $\overline{W}_{math}$, $\overline{NEL}_{math}$, $\overline{NSWD}_{math}$, $\bar{M}_{math}$, $\bar{F}_{math}$
4	$\overline{SBA}_{math}$	$\bar{E}_{math}$, ~~$\bar{I}_{math}$, $\overline{NHS}_{math}$~~, $\overline{HS}_{math}$, $\overline{BA}_{math}$, $\overline{NEL}_{math}$, $\overline{NSWD}_{math}$, $\bar{M}_{math}$, $\bar{F}_{math}$
5	$\overline{HS}_{math}$	$\bar{E}_{math}$, ~~$\bar{I}_{math}$, $\overline{NHS}_{math}$~~, $\overline{SBA}_{math}$, $\overline{BA}_{math}$, $\overline{NEL}_{math}$, $\overline{NSWD}_{math}$, $\bar{M}_{math}$, $\bar{F}_{math}$
6	$\bar{B}_{math}$	~~$\bar{E}_{math}$, $\overline{NHS}_{math}$, $\overline{HS}_{math}$, $\overline{W}_{math}$, $\bar{H}_{math}$~~
7	$\overline{NHS}_{math}$	~~$\bar{E}_{math}$, $\overline{HS}_{math}$, $\overline{SBA}_{math}$, $\overline{BA}_{math}$ $\overline{W}_{math}$, $\bar{B}_{math}$, $\bar{H}_{math}$, $\overline{AINA}_{math}$, $\overline{NEL}_{math}$~~
8	$\bar{H}_{math}$	~~$\overline{NHS}_{math}$, $\bar{B}_{math}$, $\overline{EL}_{math}$~~
9	$*\overline{AINA}_{math}$	~~$\overline{EL}_{math}$, $\overline{API}_{math}$~~
10	$\overline{EL}_{math}$	~~$\bar{H}_{math}$~~, $\overline{AINA}_{math}$

The winnowing of predictor variables that do not meet the criterion set for use of MICE in the FLEX CS approach results in estimation of mean math achievement of three of the original ten subgroups, as demonstrated in Table 3.11. For subgroups of interest that drop out of the chained equations model, a MICE subestimate does not factor into a corresponding FLEX CS estimate.

***Table 3.11*:** Visiting sequence for computing MICE subestimates in FLEX CS approach

Order	Outcome variable	Predictor variable(s)
1	$\overline{BA}_{math}$	$\bar{I}_{math}$, $\overline{HS}_{math}$, $\overline{SBA}_{math}$, $\overline{W}_{math}$, $\overline{NEL}_{math}$, $\overline{NSWD}_{math}$, $\bar{M}_{math}$, $\bar{F}_{math}$
2	$\overline{SBA}_{math}$	$\bar{E}_{math}$, $\overline{HS}_{math}$, $\overline{BA}_{math}$, $\overline{NEL}_{math}$, $\overline{NSWD}_{math}$, $\bar{M}_{math}$, $\bar{F}_{math}$
3	$\overline{HS}_{math}$	$\bar{E}_{math}$, $\overline{SBA}_{math}$, $\overline{BA}_{math}$, $\overline{NEL}_{math}$, $\overline{NSWD}_{math}$, $\bar{M}_{math}$, $\bar{F}_{math}$

2. **FH Subestimate**

An FH subestimate always factors into a corresponding FLEX CS estimate. However, calculation of the synthetic-regression estimate differs than previously described. Instead of using the full sets of predictor variables as previously described for computing estimates of mean math achievement with the FH technique, the subset of predictor variables that maximize

adjusted r-squared ($R^2{}_{adjusted}$) statistics are used for computing the regression-based estimates that contribute to the EBLUPs. Consider, for instance, the regression model previously proposed for predicting the mean math achievement of students whose parents did finish high school demonstrated in Figure 3.4.

***Figure 3.4*:** Regression model for calculating synthetic estimates of mean math achievement for students of parents who did not finish high school (NHS subgroup)

$$\widehat{\boldsymbol{Y}}_{\overline{\boldsymbol{NHS}}_{\boldsymbol{math}}} = \widehat{\boldsymbol{\beta}}_{\boldsymbol{o}} + \widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}(\%\text{B-H-AIAN}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{2}}(\text{FER}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{3}}(\%\text{EL}) + \widehat{\boldsymbol{\beta}}_{\boldsymbol{4}}(\text{SQI})$$

After fitting the outcome variable ($\overline{\boldsymbol{NHS}}_{\boldsymbol{math}}$) on all possible combinations of predictor variables from Figure 3.7, the combination that maximizes $R^2{}_{adjusted}$ is the model used for calculating the regression-based component of the FH subestimate.

3. **WPE Subestimate**

The Weighted Poststratified Estimator (WPE) is a weighted average of district-level estimates of mean math achievement of subgroups within states. The *strata* here refer to school districts within a state and the contribution of each district (i.e., stratum) to the estimate of mean math achievement of a subgroup statewide (i.e., the weighted average) is a function of the proportion of a state's subgroup population within the district. A WPE factors into a FLEX CS estimate for a particular state's subgroup if district-level estimates of mean math achievement for the state's subgroup are reported in the Stanford Education Data Archive (SEDA),[34] which includes district-level data on achievement in NAEP-referenced units (Reardon et al., 2017).

The SEDA project involves linking achievement data from mandatory standardized state assessments to the NAEP scale. NAEP-scaled estimates of achievement are available by year,

[34] For the test sample (i.e., mean math achievement of 8th graders in 2015), district-level estimates from SEDA are available in 34 of 50 states.

test subject, grade, and subgroup. However, in some instances, and for different reasons, SEDA researchers were unable to link scores for different years-subjects-grades-subgroups combinations. For this reason, WPE subestimates contribute to the FLEX CS estimate only when they are available. Calculation of the WPE for a given state's subgroup is performed through the following formula—

$$WPE = \sum_{d=1}^{D} \frac{N_d}{N}(\hat{y}_d),$$

where *D* is the number of districts for which SEDA reports estimates of mean math achievement within a state, N_d/N is the proportion of a state's subgroup population within district *d*, and $\hat{y}_d$ is a NAEP-scaled estimate of mean math achievement for district *d* per estimation by Reardon and colleagues (2017), reported in SEDA.[35]

4. **NNI Subestimate**

Nearest Neighbor Imputation (NNI) is a "donor-based" method, where an imputed value for a particular cell in a dataset comes from a value recorded for a separate but similar case in the dataset (Eskleson et al., 2009). In this study, subestimates of mean math achievement based on NNI contribute to a FLEX CS estimate if the state's nearest neighbor (i.e., most similar state) is similar enough to be considered what this study names a *sibling* state. A sibling state is defined by this study as a nearest neighbor whose Euclidean distance, a common measure of similarity, is within .40 standard deviations, where Euclidean distance is based on normalized state-level

[35] For the test sample, there are only 3 subgroups of interest (B, H, API) for which SEDA reports mean achievement by district. Further, as mentioned in the previous footnote, district achievement estimates are only available for 34 states.

measures related to academic achievement.[36] The importance of limiting use of this technique to states with siblings is that while all states have a nearest neighbor, not all pairs of neighbors are particularly similar.

Similarity of course is inherently a function of the characteristics used to draw the comparison. A pair of states may be similar in some regards but different in others. In this study, similarity can be conceived as *educational similarity*, as the characteristics used for computing a proximity matrix and determining distance between nearest neighbors represent factors known to be associated with academic achievement—including the socioeconomic and racial make-up of a state, as well as the quality of its schools (Braun & Kirsch, 2016).

The data variables used for generating Euclidean distances represent state-level measures of parental level of education, family economic resources, race and ethnicity, and school quality. Parental level of education is operationalized as the percent of states' adults 25 years or older with at least a bachelor's degree, data for which come from the American Community Survey (ACS; U.S. Census Bureau, 2018). Family economic resources is represented by a composite variable of states' median family income and net worth, data for which also come from the ACS. The race and ethnicity factor is operationalized as states' percent of the grade 8 population that identify as Black, Hispanic, American Indian, or Alaskan Native, for which data come from the Common Core of Data (CCD). School Quality is an index from Quality Counts, Education Week's annual report of states' efforts to improve public education (Education Week Research Center, 2015). Euclidean distances are commonly used to measure similarity between research

[36] A criterion value of 0.20 standard deviations was initially proposed, since this distance represents a commonly used benchmark for characterizing a difference as *small* when using standardized mean difference (SMD) to measure distances between values (Cohen, 1988; Lipsey, 2001). However, this value was doubled after observing that the minimum SMD value between pairs of states was greater than .20. See *Table 4.4* for details.

subjects, especially when comparisons are based on multiple continuous variables (Hair et al., 2009), and thus the metric lends itself to comparing states for this study.

The Euclidean distance between observations (i.e., states) is calculated from differences in values of observations on a set of variables. The Euclidean distance (d) between any pair of observations across a set of variables is given by the general formula,

$$d_{wv} = \sqrt{(x_{w1} - x_{v1})^2 + (x_{w2} - x_{v2})^2 + \cdots + (x_{wp} - x_{vp})^2}$$

Where the distance between state *w* and state *v* (d_{wv}) is equal to the square root of the sum of squared differences between observations *w* and *v*, across *p* variables $x_1, x_2, \ldots, x_p$. In the application of NNI for this dissertation, points (w, v) represent states and the Euclidean distance between states is calculated with the following formula,

$$\sqrt{(\%BA_w - \%BA_v)^2 + (FER_w - FER_v)^2 + (\%BHAINA_w - \%BHAINA_v)^2 + (SQI_w - SQI_v)^2},$$

Such that the Euclidean distance between states is equal to the square root of the sum of squared differences between states' values on measures of parental level of education (%BA), family economic resources (FER), race and ethnicity (%BHAIAN,) and school quality (SQI). To limit the undue influence of the scale on which the variables' values are measured, each of the data variables are standardized with a mean of 0 and a standard deviation of 1 so that the distribution of values for each variable are on the same scale.

Computing Final FLEX CS Estimates

Final FLEX CS estimates are precision-weighted averages of the subestimates that meet the criteria for contributing to FLEX CS estimates, much like the EBLUPs computed in the FH approach. The weights associated with subestimates are calculated in a manner that assigns greater importance to more precise subestimates. While the criteria set for using a subestimate is rather stringent, at least one of the four different types of subestimates, the FH subestimate,

always factors into each FLEX CS estimate. This subestimate only requires that a set of predictor variables that maximizes adjusted r-squared be used for computing the synthetic-regression component of the FH estimate.

For calculating each *i* of 376 FLEX CS estimates, which represent precision-weighted averages, the following general formula is used in this study—

$$FLEX\ CS = \frac{\sum_1^s (\frac{\hat{y}_{si}}{\widehat{\sigma_{s_i}}^2})}{\sum_1^s (\frac{1}{\widehat{\sigma_{s_i}}^2})},$$

where the numerator equals the sum of subestimates ($\hat{y}_s$) of *i* divided by their corresponding variance estimates $\widehat{\sigma_s}^2$ of *i* and the denominator equals the sum of 1 divided by variance estimates $\widehat{\sigma_s}^2$ of *i*. Notation *i* indexes each mean subgroup achievement value estimated with the FLEX CS technique and s_i represents the number of subestimates that contribute to each *i* FLEX CS estimate, which may vary from one to four separate subestimates.

Variance of Subestimators

Computing FLEX CS estimates requires estimation of the variance of each subestimate that meets criteria for factoring into a FLEX CS estimate. For the MICE subestimate, the variance estimate is equal to the variance of the m = 100 imputed values computed for predicting the mean math achievement of subgroups of interest. Thus, variance of MICE subestimates are equal to,

$$\frac{\sum_{i=1}^{100}(x_i - \bar{x})^2}{100 - 1}$$

Where $\bar{x}$ is the MICE estimate for a target value of interest and x_i, $\{x_1, \ldots, x_{100}\}$, represent the m imputations that are averaged in the MICE procedure for estimating the mean math achievement of a subgroup of interest.

The variance of FH subestimates (i.e., variance of EBLUPs) is a mean square error (MSE) estimator provided in the *sae* package in R through the `mseFH` command (Molina & Marhuenda, 2015). The command returns two lists. The first contains point estimates of small area means (EBLUPs), based on the Fay-Herriot model. The second contains mean square error estimates associated with each EBLUP.

The variance used for the WPE subestimate is a pooled variance, calculated as follows—

$$\frac{\sum_{i=1}^{d}(n_i - 1)\,\hat{\sigma}_i^2}{\sum_{i=1}^{d}(n_i - 1)}$$

Where d is equal to the number of districts (strata) used for estimating the WPE, n_i is equal to the number of subgroup members in district (stratum) i, and $\hat{\sigma}_i^2$ is equal to the SEDA-reported mean math achievement variance for district i.

For the NNI subestimate, the variance is equal to the variance estimate of the donor. Put differently, it is the sibling state's NAEP-reported mean variance for the corresponding subgroup. Consider, for instance, that North Dakota is the "sibling" state to South Dakota, and that AIAN students from South Dakota are a subgroup of interest. If NAEP reports an estimate of mean math achievement and variance for AIAN students in North Dakota, then the NNI subestimate factoring into the FLEX CS estimate of mean math achievement and variance for AIAN students in South Dakota is equal to the NAEP-reported estimate of mean math achievement and variance for AIAN students in North Dakota.

Application of the MICE and FH Techniques in Practice

MICE

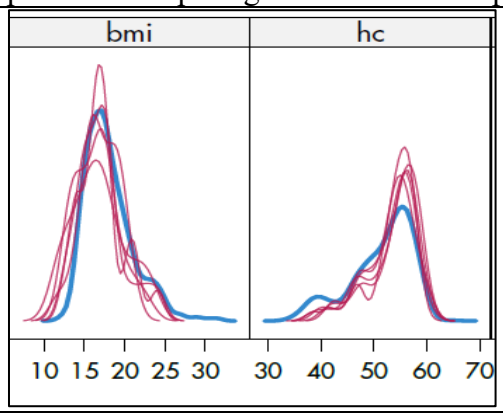
It should be noted that the application of the MICE and FH techniques described in this chapter differ in important ways from how they would be used in practice. In this study, the MICE technique involves successively removing the target values of interest from the test

sample prior to executing the MICE procedure. By contrast, removing values from an incomplete dataset is not a feature of MICE, or any imputation procedure, in practice. Removing values would only serve to reduce the effectiveness of the imputation process since observed values serve to inform which values should be imputed.

In addition, just the first of three general stages of the *mice* procedure are used for evaluating the MICE technique in this study. In practice, *mice* involves an "imputed data" stage, an "analysis results" stage, and finally a "pooled results" stage (van Buuren & Groothuis-Oudshoorn, 2011). In the first, *m* separate datasets are imputed, which results in *m* separate complete datasets. The second involves performing analyses with the *m* sets of data and the third involves pooling estimates (e.g., regression coefficients) computed across the *m* analyses.

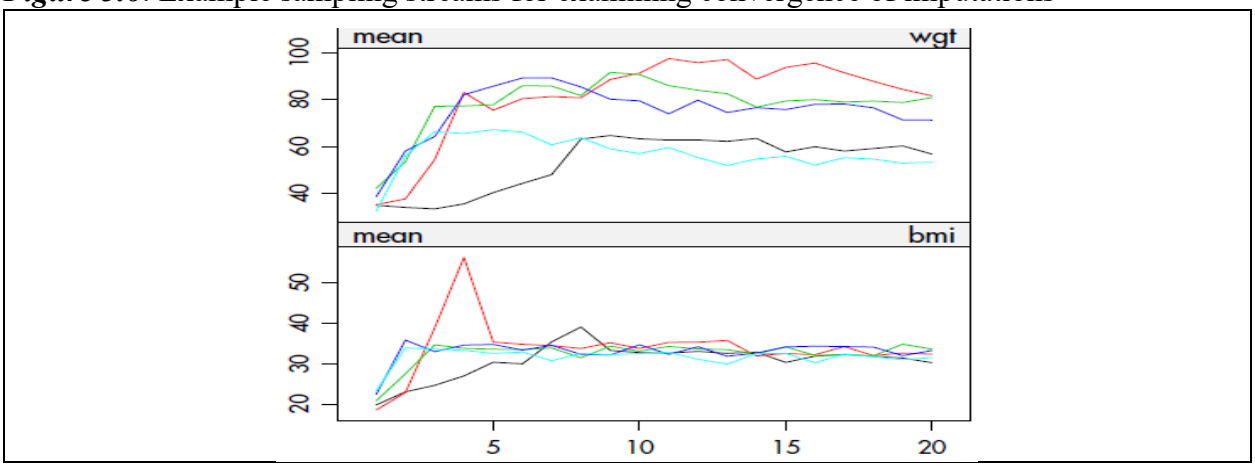
It should also be noted that diagnostic checks in *mice* for examining the plausibility of imputed values frequently involves more than just checks for out-of-range imputations proposed in this study. Additional checks, for instance, may include superimposing density plots of imputed values over observed values (such as the plots depicted in Figure 3.5) to compare their distributions. In Figure 3.5, copied from van Buuren & Groothuis-Oudshoorn (2011), the density plots for five sets of imputed values, outlined in red, for variables "bmi" and "hc" are superimposed on the density plot of observed values, outlined in blue.

***Figure 3.5*:** Example density plots for comparing distribution of imputed and observed values



Another common check is to examine the "sampling streams," which depict the various draws across iterations of the *mice* algorithm per *m* sets of imputed values. This check permits confirmation of whether imputations tend to converge around similar values by examining series of line plots, such as those depicted in Figure 3.6, copied from van Buuren & Groothuis-Oudshoorn (2011). In these example line plots, the mean of drawn values for variables "wgt" and "bmi" are plotted across iterations (t=20) for five separate sets of imputations (m=5). As can be gleaned from Figure 3.6, the mean of imputations for the "bmi" variable appear to converge around a similar value across sets of imputations, but the same degree of convergence does not occur for the "wgt" variable.

Figure 3.6: Example sampling streams for examining convergence of imputations



FH

Applying an SAE technique such as the Fay-Herriot (FH) model in practice would not involve randomly sampling from a larger previously drawn sample to compute direct estimates. The modified version of the technique described in this chapter is intended to simulate a scenario in which NAEP researchers are not able to sample enough students from a subpopulation by taking small random samples of students from state subgroups that are not actually deficient in

sample size. The approach permits a comparison between FH-generated mean subgroup estimates, based, in part, on randomly drawn samples, and mean math achievement estimates reported by NAEP (i.e., the target values). Such a design supports the evaluative nature of this study.

By contrast, in practice NAEP researchers would use all of the sampled students available to them for computing the direct estimate of any state subgroup that does not meet the rule-of-62. Thus they would use samples ranging between 0 and 61 in size, however many students are sampled who are members of the subgroup in the state. In the event no subjects are sampled for direct estimation, it is practice in Small Area Estimation to use only the synthetic (i.e., regression-based), indirect estimator (Rao, 2013).

In practice, computing direct estimates from small samples should also involve special consideration on how to protect the privacy of students. In accordance with the Federal Education Rights and Privacy Act (FERPA), education agencies are legally bound to protect students' personally identifiable information (20 U.S.C. § 1232g; 34 CFR Part 99) and publishing direct estimates from small samples risks revealing personally identifiable information. For this reason, if a form of SAE, including FH, is used in practice for estimating mean subgroup achievement, it is recommended that researchers either conceal the sample size used or abandon efforts to compute direct estimates if the available sample is too small. While FERPA requires states themselves define minimum sample size requirements for publishing achievement results, most states require that a sample represent at least 10 students (U.S. Department of Education, 2010).

Finally it should be noted that the using the FH technique in practice would not involve fitting nearly as many regression models for computing indirect (i.e., regression-synthetic)

estimates as proposed in this study. Unlike this study, which involves fitting a regression model for each target value of interest, 376 total, applying the FH model in practice would only require fitting a regression model per subgroup that includes direct estimates computed from insufficiently large samples. For the test sample (i.e., mean math achievement of 8^{th} graders in 2015), this would involve fitting 10 regression models, one per incomplete subgroup. In each of these 10 models, the outcome variable would represent a different incomplete variable from the test sample and the cells corresponding to state-subgroup pairs that are unreported (i.e., empty cells) would be filled with direct estimates computed from small samples of less than 62 students.

Chapter 4: Results

The goal of this study is to determine whether one or more of three prediction techniques are suitable for estimating subgroup performance on State NAEP. This chapter describes the results of the statistical analyses conducted to evaluate the predictive accuracy of these three techniques. It begins with a comprehensive description of the data used for each technique. Second, it describes the estimates of mean math achievement computed by each technique and contrasts these estimates with the corresponding NAEP-reported estimates (target values). Third, it presents the accuracy measures—weighted Mean Absolute Error (wMAE) and coverage—calculated for each technique. Finally, it directly addresses the dissertation's three research questions.

Description of Data Used for the MICE Technique

The data used for the MICE technique are NAEP-reported state-level estimates of mean subgroup achievement from the grade 8 NAEP math assessment in 2015 (i.e., test sample data). These data are obtained from the National Center for Education Statistics (NCES) website through the NAEP Data Explorer (NDE), a web-based system that provides users with tables of detailed results from NAEP assessments (U.S. Department of Education, 2008). These data include 18 variables with values representing mean math achievement estimates for different NAEP reporting groups (i.e., subgroups) across states. These subgroups include the 10 subgroups of interest for which the NAEP program was unable to report estimates of mean math achievement for at least one of the 50 states, and 8 subgroups for which reporting is complete (see Table 4.1).

Table 4.1: Outline of variables from the test sample used for the MICE technique

Variable (student subgroup)	Complete/Incomplete	Number of missing values
Eligible for the national free or reduced lunch program (E)	Complete	N/A
Ineligible for the national free or reduced lunch program (I)	Complete	N/A
Parents did not finish high school (NHS)	Incomplete	2
Parents graduated from high school (HS)	Incomplete	2
Parents had some education after high school (SBA)	Incomplete	2
Parents graduated from college (BA)	Incomplete	2
White (W)	Complete	N/A
Black (B)	Incomplete	11
Hispanic (H)	Incomplete	3
Asian/Pacific Islander (API)	Incomplete	20
American Indian/Alaskan Native (AIAN)	Incomplete	37
Two or more races (TP)	Incomplete	26
English language learner (EL)	Incomplete	19
Not an English language learner (NEL)	Complete	N/A
Student with disability (SWD)	Complete	N/A
Not a student with a disability (NSWD)	Complete	N/A
Male (M)	Complete	N/A
Female (F)	Complete	N/A

Description of Data Used for the FH Technique

The data used for computing estimates of mean math achievement with the FH technique come from restricted-use student-level data from the National Center for Education Statistics (U.S. Department of Education, 2020c) and public-use state-level data from the Common Core of Data (2020a), Education Week (Education Week Research Center, 2015) and the U.S. Census Bureau (2018). The student data from NCES are the plausible values of grade 8 math achievement from NAEP testing in 2015. The administrative data represent state-level factors (characteristics) related to academic achievement. The student-level data are used to calculate direct estimates of mean math achievement for each state-subgroup pair of interest.[37] In the FH

[37] Direct estimates are computed that account for NAEP's complex sampling design. As demonstrated in Appendix A in the section titled "Computing Direct Estimates," the utilized code calls commands that instruct the software program (STATA) to use each student's sampling weight (*ORIGWT*), jackknife replicate weights for their cluster (*SRWT'*s), and plausible values of grade 8 math achievement (*MRPCM*'s).

technique, direct estimates are combined, by subgroup of interest, with regression-based estimates of mean math achievement calculated by fitting direct estimates on predictor variables created from administrative data.[38] An overview of these predictor variables, constructed from administrative data sets, are presented in Table 4.2. This overview includes the predictor variables' names, the factors they are intended to represent, their operational definition, the subgroups whose mean math achievement they predict, and their source of data.

Table 4.2: Outline of predictor variables used for the FH technique

Name & Factor	Operational definition	Subgroup(s) whose mean achievement is predicted	Data source
%B-H-AIAN, Race/Ethnicity of students	State percent of grade 8 students who identify as Black, Hispanic, American Indian, or Alaskan Native	NHS, HS, SBA, BA, TP, EL	NCES (CCD)
FER, Economic circumstances of students' families	The mean of a state's median household income and wealth in dollars	NHS, HS, SBA, BA, B, H, API, AIAN, TP, EL	U.S. Census
%EL, English proficiency of students	State percent of students identified as English learners	NHS, HS, SBA, BA, H, API	NCES (CCD)
SQI, School Quality	An indicator of school quality in each state measured on a continuous scale ranging from 0 to 100	NHS, HS, SBA, BA, B, H, API, TP, EL	Education Week
%BA, Parental level of education	State percent of adults 25 years or older that have earned a bachelor's or more advanced degree	B, H, API, AIAN, TP	U.S. Census
%AA, Black ethnicity	State percent of Black population born in the US	B	U.S. Census
%MX, Hispanic origin	State percent of Hispanic population of Mexican descent.	H	U.S. Census
%A, Asian background	State percent of grade 8 students who identify as Asian, but not Pacific Islander	API	NCES (CCD)

[38] The response variable values all represent direct estimates of some form. For each regression model fit for the FH approach (376 total), one value from the response variable is a direct estimate based on a small sample ($n < 62$) randomly drawn from the restricted-use data and the rest of the response variable values are NAEP-reported direct estimates. This iterative maneuver, by which NAEP-reported direct estimates are successively replaced with small sample direct estimates, is used to simulate the sort of sample and estimate reliability researchers would have in practice.

It should be noted that most of the administrative data used for constructing predictor variables for the FH technique are indirect measures (i.e., proxy measures) of the state-level factors they are meant to represent. The data of interest in this study represent the *achievement of grade 8 students in 2015 in public schools*. However, the sources from which various administrative data are collected do not provide data on this group of students specifically.[39] The discrepancies between the predictor variables and the factors they are intended to represent are discussed in the following paragraphs.

Although the *%B_H_AIAN* variable represents the proportion of grade 8 Black, Hispanic and American Indian or Alaskan Native students across states in 2015, it reflects both public and private school students. The source of these data, The Common Core of Data (CCD), does not separate public and private school students at the subgroup level across states.

There are two notable limitations with the *FER* variable. First, available administrative data on household income and wealth are not disaggregated by households with children of different ages (e.g., grade 8 age students), nor are they disaggregated by whether there are children living in a household. Consequently, the *FER* variable is a measure of economic resources of all types of households across states. Second, the data used to represent state-to-state variation in families' wealth reflect 2013 median household estimates, which are available through a special study conducted by the U.S. Census Bureau (Cheneverth et al., 2017).

[39] While it is possible to calculate *some* of these predictor variables used in this analysis directly from restricted-use data, this particular strategy is not pursued for reliability concerns. There are at least a couple of issues that would negatively impact the reliability of predictor variables computed from restricted-use data. The main issue is related to small samples— estimates for certain states would be based on very small samples or they would be impossible to compute because they do not exist in the restricted-use data. Consider, for instance, the manner by which the %A variable is constructed – proportion of API that identifies as "A" (Asian, not Pacific Islander). There may not be *any* PI students sampled by NAEP in certain states. Another issue, which threatens to contribute further measurement error to estimates, relates to self-reporting with children. There is some research evidence indicating that students often misreport their parents' level of education (Kreuter et al., 2010).

Unfortunately, these special studies are not conducted biannually, and a similar study was not conducted for 2015.

The *%EL* variable does represent the percent of English learners across states in a specific grade. Although the Common Core of Data includes data on English learners disaggregated by grade level nationally, disaggregation by grade level is not available for individual states. Therefore, the %EL variable is a measure of the percent of English learners by states in grades K-12.

The *%AA* variable does not measure the proportion of Black students across states who identify as African-American. This level of detail is not available in U.S. Census data. Instead, this variable represents state-level estimates of the proportion of the Black population that is born in the United States. Similarly, the %MX variable does not represent the proportion of Hispanic students of Mexican origin across states. Again, this level of disaggregation by grade or age does not exist in U.S. Census data. Instead, this variable represents an estimate of the proportion of the Hispanic population of Mexican origin of all ages in each state.

Description of Data Used for the FLEX CS Technique

The data used in the FLEX CS technique include the data used in the MICE and FH techniques. In addition, data from the Stanford Education Data Archive (SEDA, Reardon et al., 2017) are used to compute weighted poststratified estimators (WPEs). It should be noted, however, that the opportunity to use the WPEs in the calculation of FLEX CS estimates is limited inasmuch as the SEDA data available for the test sample (i.e., grade 8 Math in 2015) include NAEP-referenced achievement data for just 3 subgroups of interest and 34 states. Table

4.3 displays the subgroups and corresponding states for which NAEP-referenced achievement data from SEDA are available for the test sample.

Table 4.3: Subgroups and states for which NAEP-referenced achievement data are available in SEDA for grade 8 math in 2015

Subgroup	States
Black (B)	AL, AK, AZ, CA, CT, GA, HI, ID, IN, IA, KS, KY, LA, MD, MA, MI, MN, MS, NE, NM, NC, OK, PA, SC, SD, TN, UT, WV, WY (30)
Hispanic (H)	AL, AK, AZ, CA, CT, DE, FL, GA, HI, ID, IN, IA, KS, KY, LA, MD, MI, MN, MS, NE, NM, NC, OK, OR, PA, SC, SD, TN, UT, WV, WI, WY (33)
Asian Pacific Islander (API)	AL, AK, AZ, CA, CT, DE, FL, GA, HI, ID, IL, IN, IA, KS, KY, LA, MD, MA, MI, MN, MS, NE, NM, NC, OK, OR, PA, SC, SD, UT, WV, WI (32)

An additional type of subestimate used for calculating FLEX CS estimates of mean math achievement is a Nearest Neighbor Imputation (NNI) subestimate. The NNI subestimate is a donor-based estimate, where the estimate of mean math achievement for one state takes on the observed value (i.e., NAEP-reported estimate of mean math achievement) of its "nearest neighbor," meaning the state with which it is most similar based on a select set of characteristics. The data used for the NNI subestimate include data used in the FH technique. Specifically, variables representing parental level of education (*%BA*), family economic resources (*FER*), race and ethnicity *(%B-H-AIAN*), and school quality (*SQI*). The *%BA* and *FER* variables are constructed from American Community Survey data, the *%B-H-AIAN* variable is constructed from National Center for Educational Statistics data and the *SQI* variable is a measure created by Education Week.

The criterion established for determining whether an NNI subestimate contributes to a FLEX CS estimate depended on whether pairs of states were similar enough to be considered *sibling* states—originally defined as states whose Euclidean distance (based on the *%BA*, *FER*, *%B-H-AIAN*, and *SQI* variables) was less than 0.20. This criterion resulted too stringent, as not a single pair of states shared a distance of less than 0.20. To accommodate use of NNI

subestimates, a less stringent criterion of 0.40 was used (double the proposed distance). Redefining *sibling* states as pairs of states whose Euclidean distance was less than 0.40 resulted in using NNI subestimates in the calculation of FLEX CS estimates of mean math achievement for 12 states (6 pairs), presented in Table 4.4.

Table 4.4: "Sibling states," pairs of states whose Euclidean distance is less than 0.40

State pairs	Euclidean distance
Alabama (AL) & Oklahoma (OK)	0.25
Connecticut (CT) & New Jersey (NJ)	0.36
Iowa (IA) & North Dakota (ND)	0.39
Kansas (KS) & Nebraska (NE)	0.36
Michigan (MI) & Missouri (MO)	0.28
Pennsylvania (PA) & Wisconsin (WI)	0.25

Note: Euclidean distance is calculated with continuous variables representing four separate state-level factors related to parental level of education, families' economic circumstances, race & ethnicity, and quality of schools.

Estimates of Mean Math Achievement with the MICE technique

This section begins with a discussion of results from diagnostic checks performed, by subgroup of interest, on *preliminary* sets of MICE-produced estimates of mean math achievement. These initial results determined whether *final* sets of MICE-produced estimates would be calculated through the normal linear regression model or through Predictive Mean Matching (PMM). Then, per subgroup of interest, this section discusses and contrasts the distribution of estimates produced with the MICE technique against corresponding NAEP-reported estimates—the target values. Boxplots and histograms of the distributions are presented to support the comparison of estimates. Finally, a summary table with descriptive statistics of MICE-produced and NAEP-reported estimates is presented with remarks summarizing the extent to which the MICE technique appears to predict NAEP-reported estimates of mean math achievement.

Results from Diagnostics of Averaged Imputations

As proposed in chapter 3, diagnostic checks were conducted to evaluate the plausibility of preliminary sets of predicted values with the MICE technique per subgroup of interest. The intention was to assess whether the predicted values fell within a reasonable range of expected values by subgroup. If these values fell outside subgroup-specific intervals deemed to span a range of credible values, then predicted values for the subgroup were recalculated with Predictive Mean Matching (PMM), rather than the normal linear regression model with the *mice* algorithm. The range of credible values were calculated per subgroup using Tukey's (1977) "$1.5 \times IQR$" (inter-quartile range) rule for detecting outlying observations with NAEP-reported estimates of mean math achievement. Outlying observations, per Tukey's rule, were deemed to fall outside of a range of credible values.

Results from these checks indicated that for three of the ten subgroups of interest, initial estimates of mean math achievement fell outside of their pre-specified ranges of credible values when calculated through the normal linear regression model. These results, illustrated in Figure C.1.1 in Appendix C, demonstrate out-of-bound mean math achievement estimates for two parental level of education subgroups, students whose parents experienced some college (SBA) and students whose parents earned at least a bachelor's degree (BA), and for one race and ethnicity subgroup, students identifying as American Indian or Alaskan Native (AIAN).

For each of these three subgroups, one predicted value of mean math achievement was greater than its pre-specified upper bound of credible values and one value was less than its lower bound. For the parental level of education subgroups (SBA and BA), the mean achievement estimate of students from Massachusetts was the higher out-of-bound estimate and the mean achievement estimate of students from Alabama was the lower out-of-bound estimate.

For the AIAN subgroup, the mean achievement estimate for students from Minnesota was the higher out-of-bound estimate and the estimate for students from Arizona was the lower out-of-bound estimate. Accordingly, MICE-produced estimates of mean math achievement for the SBA, BA, and AIAN subgroups were recalculated with Predictive Mean Matching (PMM). Diagnostics conducted on these newly computed estimates (see Figure C.1.2 in Appendix C), demonstrate that these PMM-based estimates fell within their corresponding ranges of credible values.

Description of MICE Estimates by Subgroup

The mean and median values of NAEP-reported and MICE-produced estimates of mean math achievement are similar across subgroups. Rounded to the nearest integer,[40] the mean values of NAEP and MICE estimates are equal for 9 of 10 subgroups. For the subgroup representing American Indian and Alaskan Native (AIAN) students, the mean of MICE estimates is one point greater (NAEP mean = 259, MICE mean = 260).

The median values of NAEP and MICE estimates are equal for 7 of 10 subgroups. These median values are one point apart for the NHS (NAEP median = 266, MICE median = 265) and HS (NAEP median = 268, MICE median = 269) subgroups. For the subgroup representing Black students (B), the median values are two points apart (NAEP median = 258, MICE median = 260).

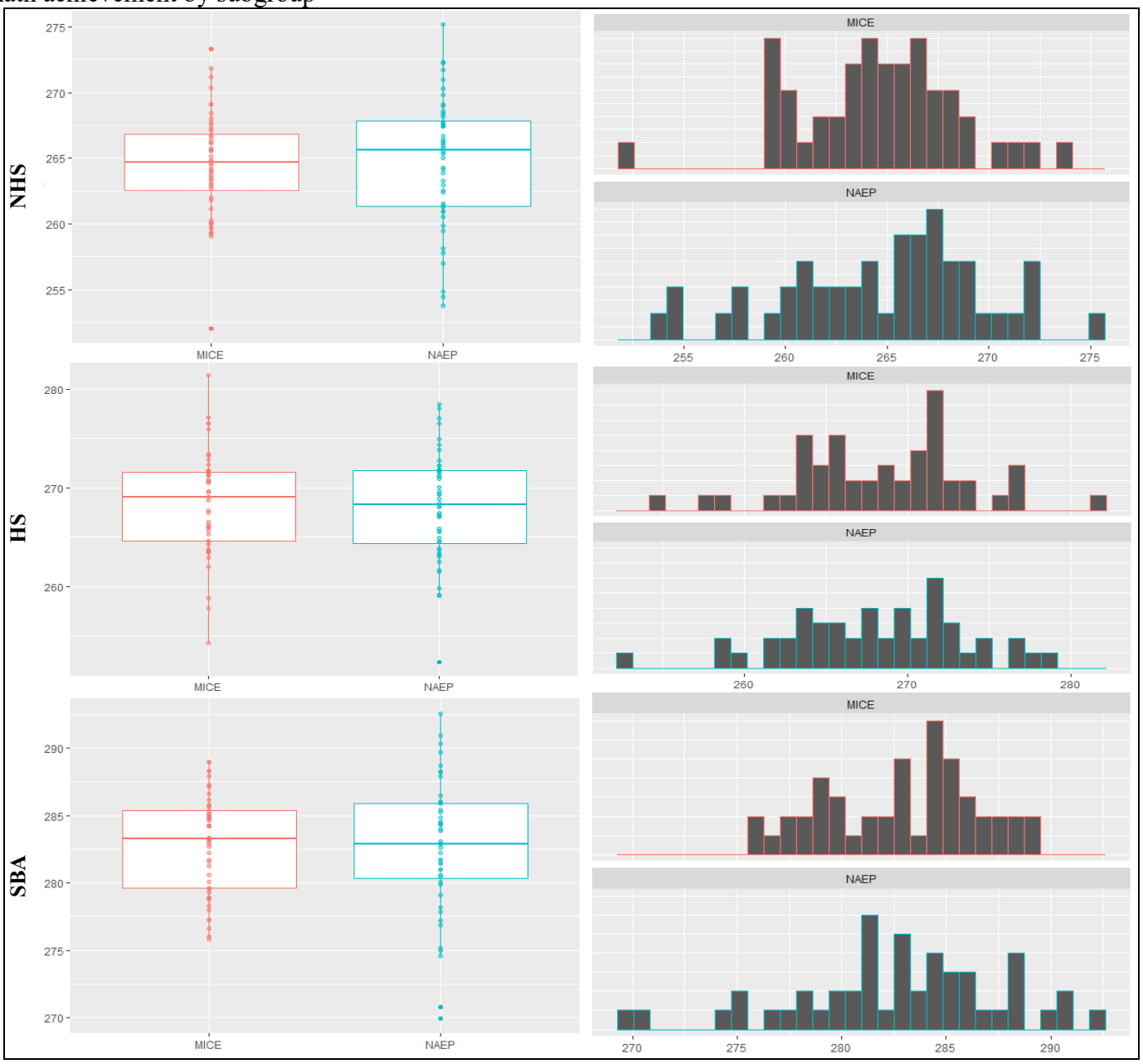
For all subgroups of interest, NAEP-reported estimates of mean math achievement are more variable than MICE-produced estimates, a trend that can be observed through the series of boxplots and histograms in Figure 4.1.1. In other words, the standard deviations of NAEP estimates are larger than corresponding standard deviations of MICE estimates. These

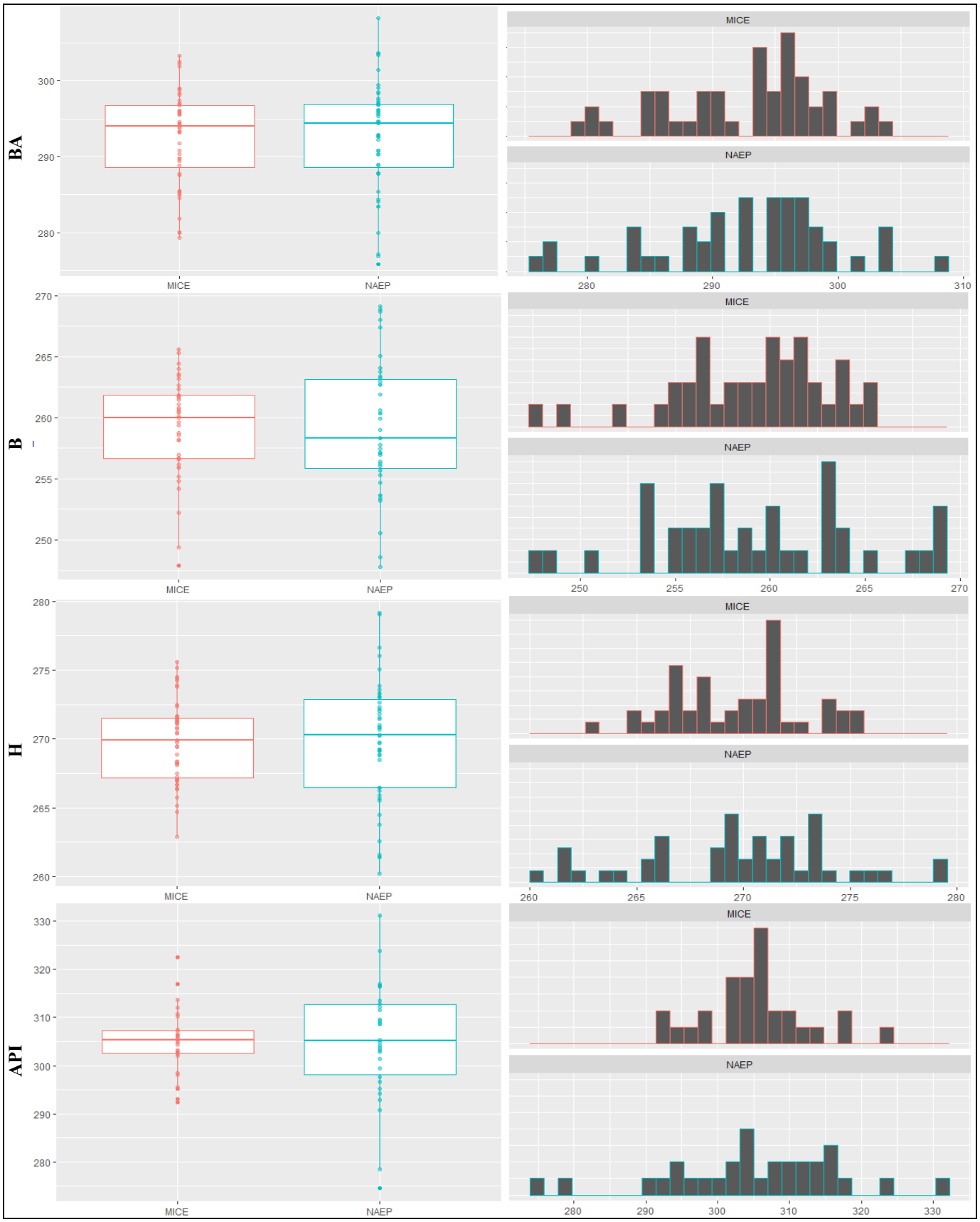
[40] All estimates of mean achievement, per NAEP reporting convention, are rounded to their nearest integer.

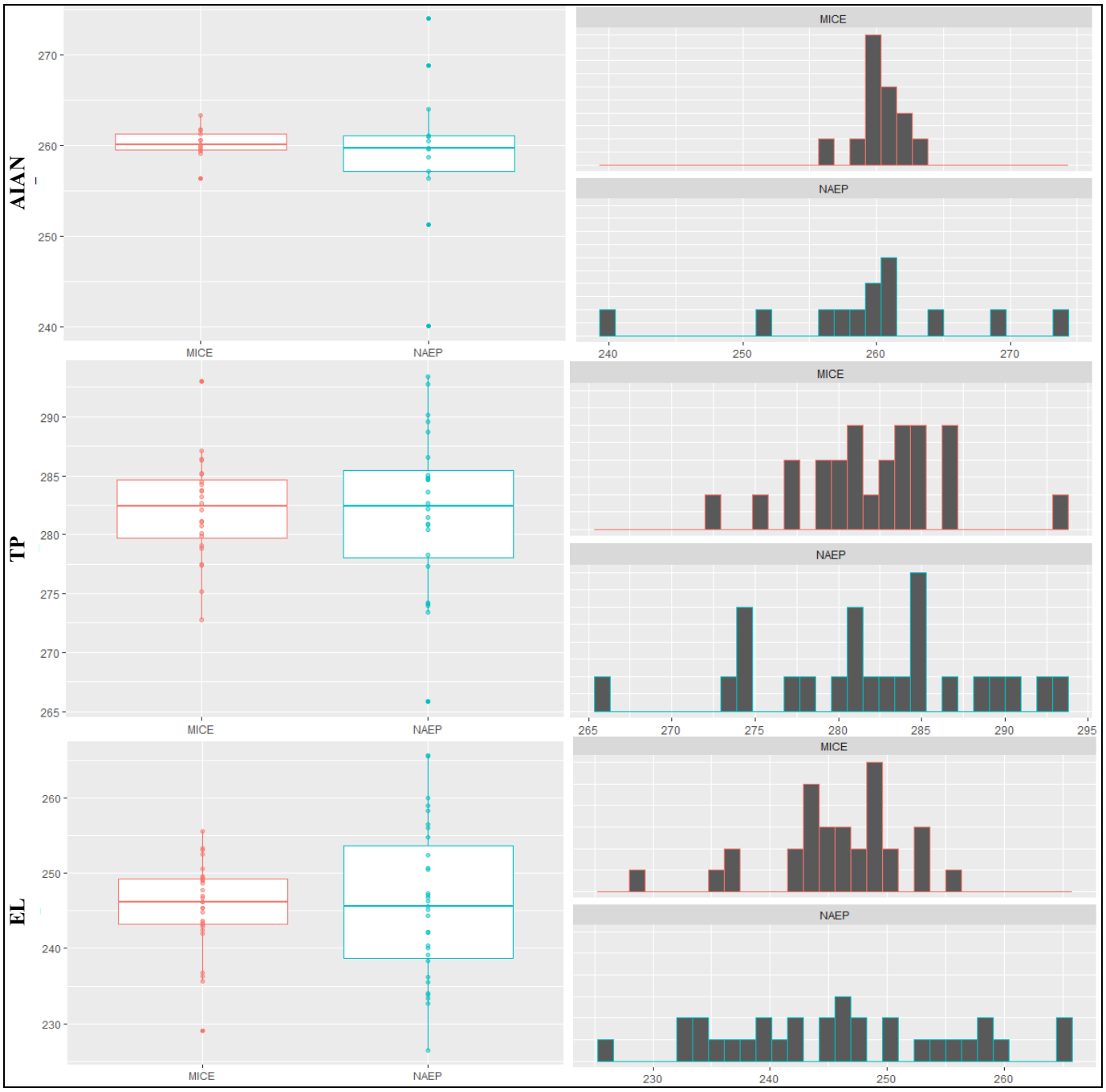
discrepancies in variance are especially pronounced for the subgroups representing Asian Pacific Islander (API) and American Indian and Alaskan Native (AIAN) students. For the API subgroup, for instance, NAEP-reported estimates of mean achievement range from 275 to 331 (56 points), while MICE-produced estimates for this subgroup range from 292 to 323 (31 points). For the AIAN subgroup, NAEP estimates range from 240 to 274 (34 points), while MICE estimates range from 256 to 263 (7 points).

It should also be noted that, among the parental level of education subgroups, the sets of MICE estimates calculated with Predictive Mean Matching (PMM) are far less variable than corresponding NAEP estimates, compared to sets of MICE estimates calculated through normal linear regression. For the SBA and BA subgroups, for which MICE estimates were produced through PMM, the difference in the range of NAEP and MICE estimates are 10 and 9 points, respectively. For the NHS and HS subgroups, for which NAEP estimates were produced through normal linear regression, the range in sets of values is equal for the former subgroup and different by one point in the latter. A full account of descriptive statistics of NAEP-reported vs. MICE-produced estimates of mean math achievement by subgroup is presented in Table 4.1.5.

Figure 4.1.1: Boxplots and histograms of NAEP-reported vs MICE-produced estimates of mean math achievement by subgroup




BA
B
H
API
MICE
NAEP
300
290
280
270
265
260
255
250
275
330
320
310



For two subgroups, NHS and HS, the MICE technique successfully produces estimates of mean math achievement that approximate target values (i.e., NAEP estimates) lying at the lower tail of their respective distributions. For the HS subgroup, for example, the minimum NAEP-reported estimate is an outlying value equal to 252, for students in Alabama, which is seven

points lower than the second lowest NAEP estimate. However, the MICE-produced estimate of mean achievement for Alabama is only two points greater (254), a similarity best appreciated by reviewing the boxplot for the NHS subgroup in Figure 4.1.1.

For other subgroups, the MICE technique is unable to produce estimates of mean math achievement that approximate target values lying along the tails of their respective distributions. A comparison of MICE and NAEP estimates for the TP and EL subgroups serve as helpful examples. The minimum NAEP estimate for the TP subgroup, for students in Kentucky, equals 266. Meanwhile the next lowest-achieving state, per NAEP, for this subgroup is Oklahoma with a mean achievement estimate equal to 273. The minimum MICE estimate for this TP subgroup equals 273. For the EL subgroup, the maximum NAEP-reported estimate of mean math achievement equals 266, for students in both Kentucky and South Carolina. The maximum MICE estimate of mean math achievement for this subgroup, for students in Alaska, equals 256—a full 10 points less.

Summary Remarks on Descriptive Statistics of MICE Estimates

The MICE technique tends to produce sets of estimates that cluster around the center of corresponding distributions of target values. The MICE estimates therefore appear to be generally biased toward the state averages of mean achievement across subgroups. This truncating effect is particularly problematic when predicted values are calculated with Predictive Mean Matching (PMM). This last point should not be too surprising considering imputed values with PMM are by definition constrained to equal an *observed* value selected at random. Accordingly, it is not possible for an imputed value (i.e., predicted value, in this case) to be greater or less than maximum or minimum observed values, even when an estimand's true value is less or greater than any of the observed values.

Table 4.1.1: Descriptive statistics of NAEP-reported vs. MICE-produced estimates of mean math achievement by subgroup of interest

	Mean		**SD**		**Min**		**Median**		**Max**		**Range**	
	NAEP	MICE	NAEP	MICE	NAEP	MICE	NAEP	MICE	NAEP	MICE	NAEP	MICE
NHS	265	265	5	4	254	252	266	265	275	273	21	21
HS	268	268	6	5	252	254	268	269	278	281	26	27
SBA	283	283	5	4	270	276	283	283	293	289	23	13
BA	293	293	7	6	276	280	294	294	308	303	32	23
B	259	259	5	4	248	248	258	260	269	266	24	21
H	270	270	4	3	260	263	270	270	279	276	19	13
API	305	305	12	7	275	292	305	305	331	323	56	31
AIAN	259	260	8	2	240	256	260	260	274	263	34	7
TP	282	282	7	4	266	273	282	282	293	293	27	20
EL	246	246	10	6	226	229	246	246	266	256	40	27

Note: Values are rounded to their nearest integer to align with NAEP-reporting convention.

The marked discrepancy in the variability of NAEP-reported and MICE-produced estimates of mean math achievement provides an early indication that the MICE technique may not be particularly effective at predicting the mean math achievement of relatively low- and high-performing states across subgroups (i.e., states whose mean achievement are near the lower and upper tails of their respective distributions). This discrepancy in the spread of NAEP and MICE estimates is particularly acute for the API and AIAN subgroups, which have relatively variable NAEP-reported estimates of mean math achievement.

Accuracy Statistics for the MICE Technique

The overall weighted Mean Absolute Error (wMAE) across subgroups of interest for the MICE technique is 1.30. The wMAE statistic is smallest (most accurate) for students whose parents' highest level of education is high school (HS; wMAE = 0.85) and largest (least accurate) for Asian Pacific Islander students (API; wMAE = 2.73). It should also be noted that the target values (NAEP-reported estimates) for the API subgroup are the most variable. The standard deviation of NAEP-reported estimates of mean math achievement for the API subgroup is 12. By contrast, the standard deviations of the target values for the other subgroups range from 4 to 10.

Table 4.1.2 presents the accuracy statistics for the MICE technique by subgroup, as well as across subgroups.

Table 4.1.2: Accuracy statistics for the MICE technique by subgroup

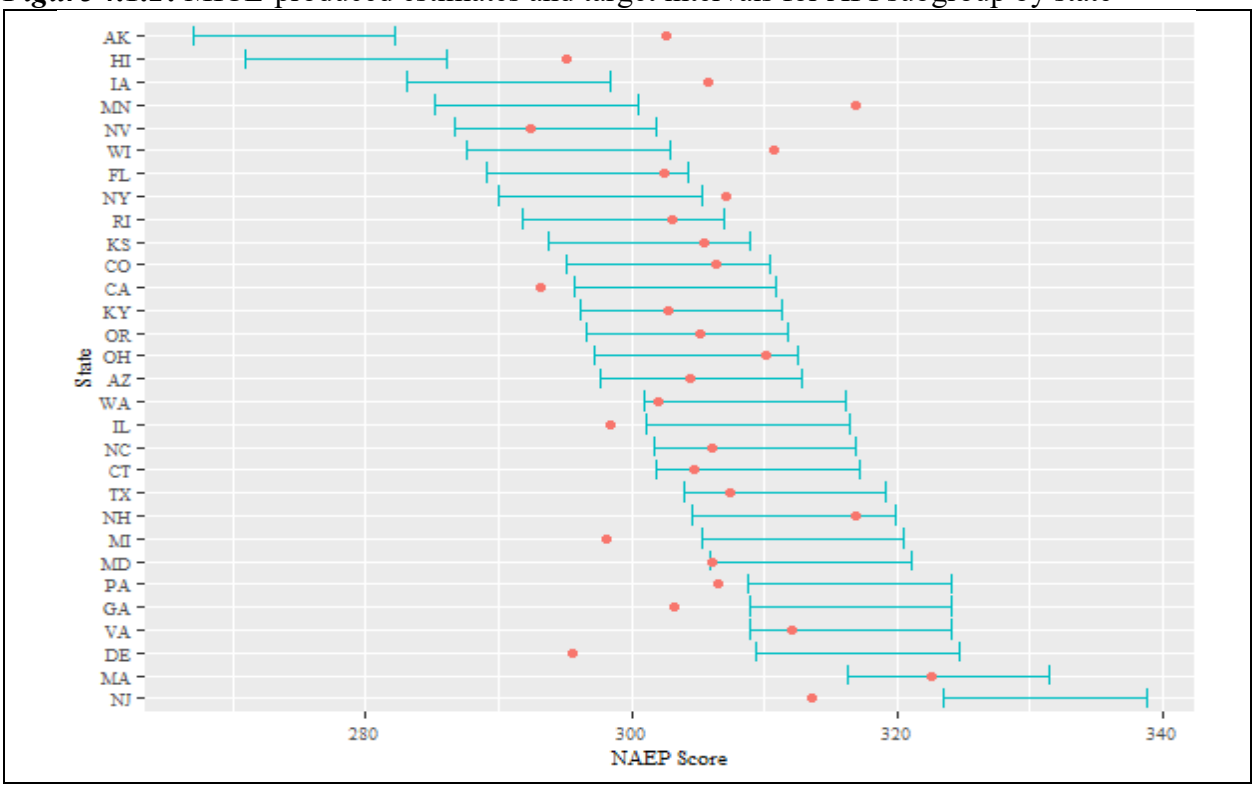
Subgroups	weighted Mean Absolute Error (wMAE)	Coverage
Did not finish high school; NHS (n = 48)	1.01	.92
Graduated high school; HS (n = 48)	.85	1.00
Some education after high school; SBA (n = 48)*	1.05	.98
Graduated from college; BA (n = 48)*	1.16	1.00
Black; B (n = 39)	1.35	.90
Hispanic; H (n = 47)	1.23	.98
Asian/Pacific Islander; API (n = 30)	2.73	.57
American Indian/Alaskan Native; AIAN (n = 13)*	1.41	.69
Two or more races; TP (n = 24)	1.09	.67
English learner; EL (n = 31)	1.85	.65
Overall (n = 376)	1.30	.88

Note: *Estimates of mean achievement computed with Predictive Mean Matching (PMM).

The overall coverage statistic across subgroups of interest for the MICE technique is 0.88, meaning 88 percent of MICE estimates of mean math achievement fall within their respective target intervals. The coverage statistics are particularly high for the parental level of education subgroups, ranging from .92 to 1.00. These statistics, however, are much lower for the race and ethnicity subgroups. The coverage statistics for the API, AIAN and TP subgroups are .57, .69 and .67, respectively. The statistic is also relatively low for the English learner (EL) subgroup, .65. For visual representations of the MICE-produced estimates of mean math achievement that “hit” and “miss” their corresponding target intervals by subgroup, see

Appendix C. For the sake of demonstration, this graph for the API subgroup, for which MICE is least accurate, is presented below.

***Figure 4.1.2*:** MICE-produced estimates and target intervals for API subgroup by state



Note: This figure reappears as *Figure C.1.9* in Appendix C.

As is evident in Figure 4.1.2, the MICE technique does not accurately predict the mean achievement of API students from lower- and higher-performing states. The MICE estimates of mean math achievement, represented by the dots, cluster near the center of the range of NAEP-reported estimates of mean math achievement. Note that the MICE predictions miss the target intervals of the four lowest-performing states for this subgroup: Alaska, Hawaii, Iowa, and Minnesota. The target intervals for these states, whose lower and upper bounds are defined as

values 0.20 standard deviations below and above their NAEP-reported estimates, cover ranges that are markedly lower than the MICE-produced estimates (i.e., predictions) for these states.

Estimates of Mean Math Achievement with the FH technique

This section begins with a brief discussion regarding the size of samples used for computing direct estimates, which are combined with regression estimates in the FH technique to form precision-weighted estimates of mean math achievement (i.e., EBLUPs). As discussed in the previous chapter, direct estimates are calculated with randomly drawn samples varying in size by subgroup of interest. Specifically, the sample sizes are equal to the median sample size of students available from the restricted-use data for the respective subgroup of interest in states that *do not* meet the rule-of-62. This permits the simulation of scenarios in which researchers have small samples of students ($n < 62$) from which to compute direct estimates of mean math achievement.[41]

Table 4.2.1 demonstrates the sample sizes used for computing direct estimates in the FH approach by subgroup, rounded here to nearest 10 to comply with National Center for Education Statistics reporting policies (U.S. Department of Education, 2020c).[42] The sample sizes range from about 10, for the subgroup representing American Indian and Alaskan Native students (AIAN) to about 50 students for six of the ten separate subgroups.[43]

[41] The Stata code used for determining sample can be found on the author's GitHub page within the section titled "Determining Sample Sizes to Draw."

[42] NCES requires sample counts from analysis with restricted-used data (RUD) to be reported to the nearest 10 to protect the privacy of students. Rounding sample sizes in this manner makes it more difficult for "data snoopers" to use these counts along with other publications based on the RUD to disclose the identity of sample respondents (U.S. Department of Education, 2020c).

[43] For grade 8 math in 2015 (the test sample), the two states—Alaska (AK) and Utah (UT)—that did not meet the rule-of-62 for the parental level of education subgroups—NHS, HS, SBA, and BA—did not report the parental level of education of *any* tested student. Thus, there was no median sample size to draw for these subgroups. Instead, for these parental level of education subgroups, samples equal in size to the largest sample used for computing direct estimates were used (about 50, which are the sample sizes used for computing direct estimates for the Black (B) and English learner (EL) subgroups. The decision to use a relatively large sample (about 50 students) is reasonable since the NAEP program is *typically* able to report the mean achievement estimates for these subgroups for *all* 50 states.

Table 4.2.1: Samples sizes used for computing direct estimates

Subgroup	Sample size (rounded to nearest 10)
NHS	50
HS	50
SBA	50
BA	50
B	50
H	40
API	40
AIAN	10
TP	30
EL	50

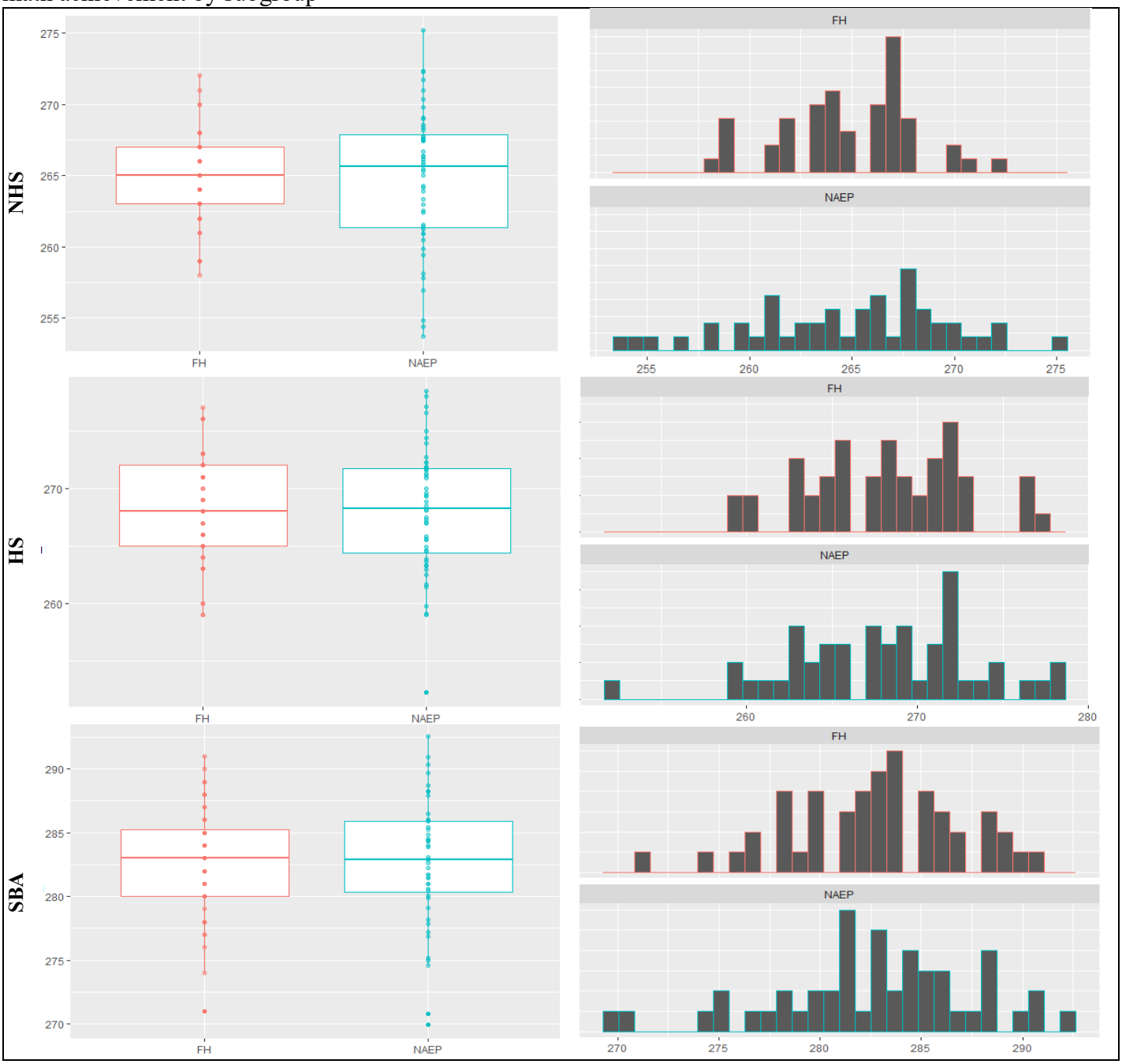
Description of FH Estimates by Subgroup

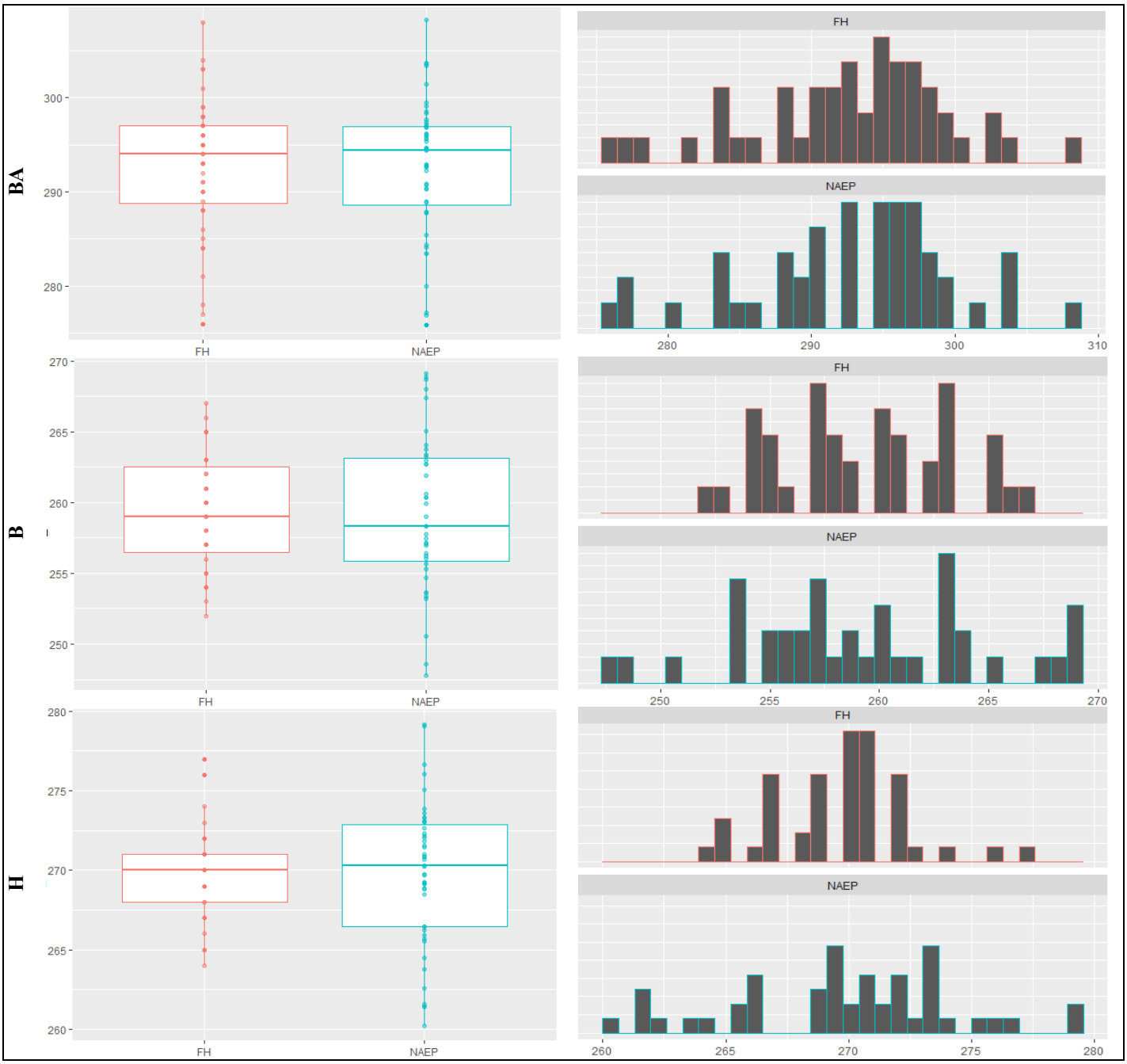
The mean and median values of NAEP-reported and FH-produced estimates of mean math achievement are even more similar across subgroups compared to NAEP and MICE estimates. The mean values of NAEP and FH estimates are equal for each subgroup. By contrast, mean estimates were equal for 9 of 10 subgroups with the MICE technique. As for sets of NAEP and MICE estimates, the median values of NAEP and FH estimates are equal for 7 of 10 subgroups. For these remaining 3 subgroups, NAEP and FH estimates are just one point apart. By contrast, the median of NAEP and MICE estimates were two points apart for the subgroup representing Black students (B). The subgroups for which the median value of NAEP and FH estimates differ by one point include the NHS (NAEP median = 266, FH median = 265), B (NAEP median = 258, FH median = 259) and AIAN (NAEP median = 259, FH median = 260) subgroups.

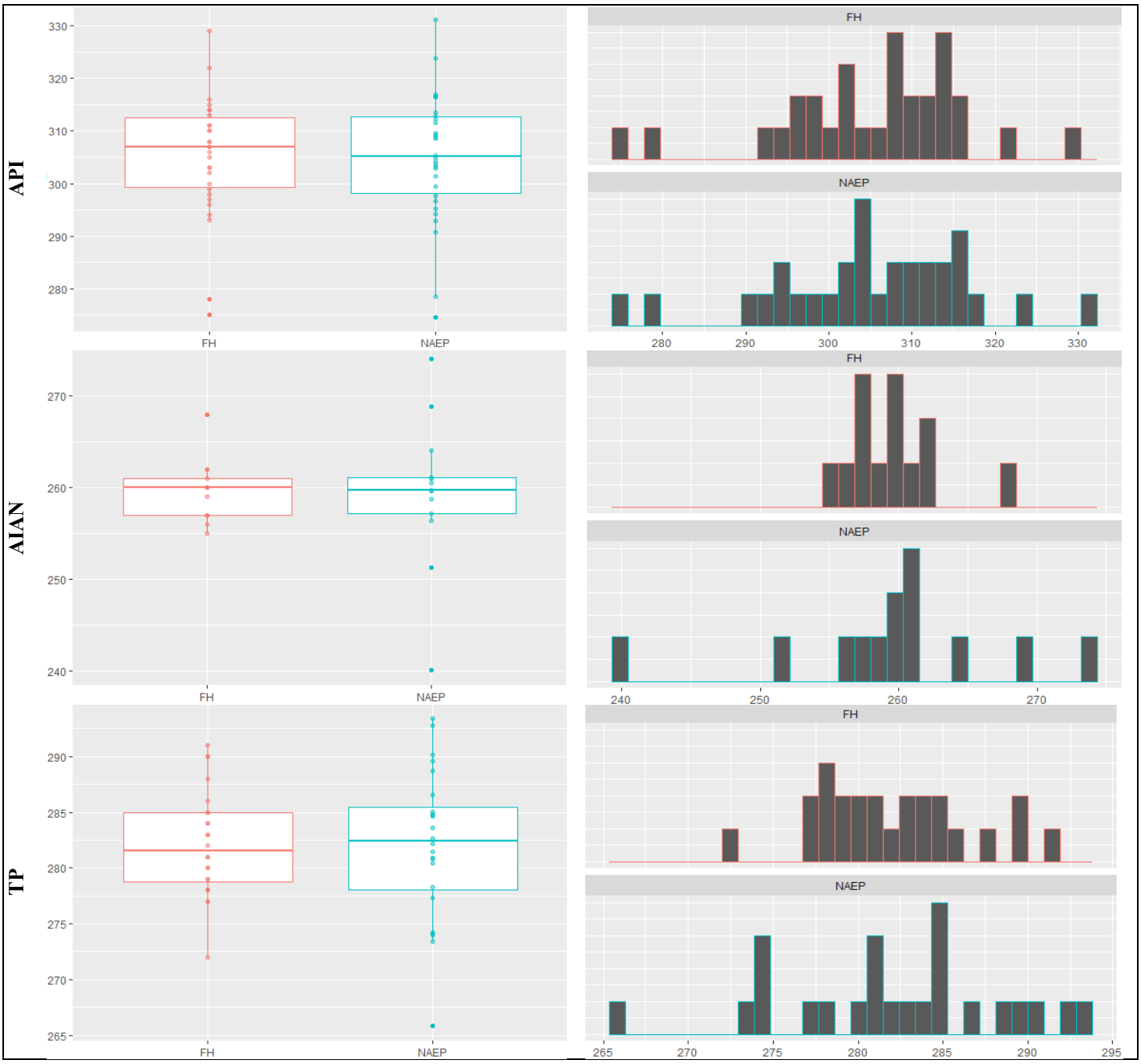
For most of the subgroups of interest, NAEP-reported estimates of mean math achievement are more variable than FH-produced estimates (see Figure 4.2.1). For one subgroup, representing students whose parents graduated from college (BA), the standard deviation of NAEP and FH estimate values are equal. The difference in the variance of estimate values is greatest for the AIAN subgroup, though this discrepancy is considerably less extreme than the

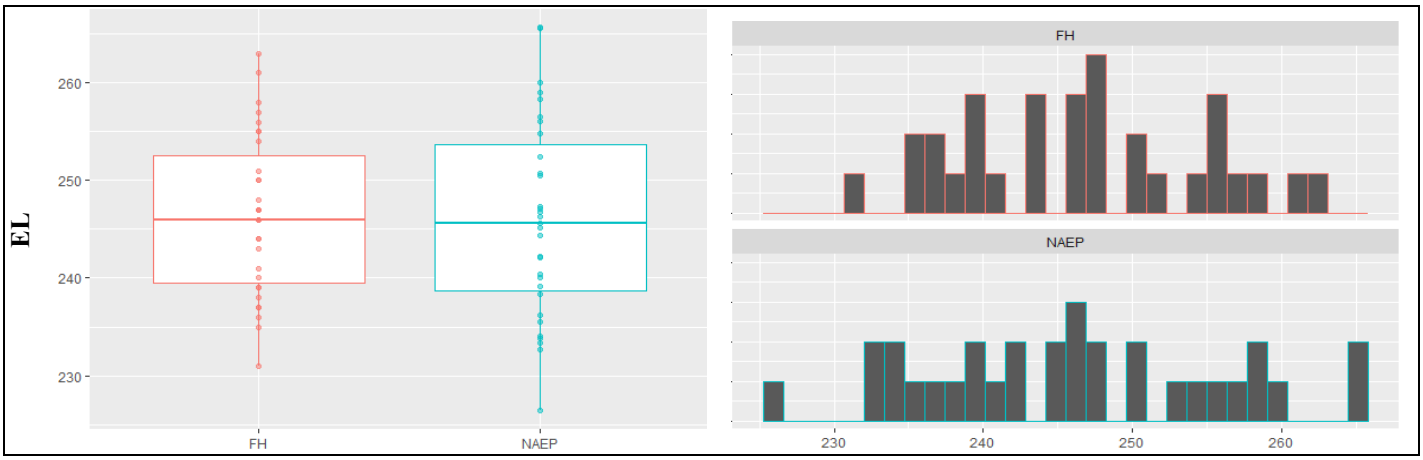
difference in the variances of NAEP and *MICE* estimate values for this subgroup. NAEP estimates for the AIAN subgroup range from 240 to 274 (34 points), while FH estimates range from 255 to 268 (13 points). By contrast, the MICE-produced estimates for this subgroup ranged just 7 points. A more complete account of the similarities and differences between sets of NAEP and FH estimates is provided in Table 4.2.2.

Figure 4.2.1: Boxplots and histograms of NAEP-reported vs FH-produced estimates of mean math achievement by subgroup




BA
B
H
FH
NAEP
300
290
280
270
265
260
255
250
275
310


API
AIAN
TP
FH
NAEP
330
320
310
300
290
280
270
260
250
240
285
275
265
295



The similarity in the variance of NAEP and FH estimates for the API subgroup is noteworthy considering the range of NAEP estimates for the API subgroup is much greater than the range of estimates of any other subgroup. NAEP estimates for the API subgroup range from 275 to 331 (56 points). By comparison, the second largest range of NAEP-reported estimates for any other subgroup is 40, for the EL subgroup. The FH estimates for the API subgroup span 54 points, from 275 to 329, which is two points less than the range of NAEP estimates. By contrast, the difference in range between NAEP-reported and *MICE*-produced estimates of mean achievement for the API subgroup was 25 points.

For a handful of other subgroups, the FH technique is less successful at producing estimate values covering the range of their corresponding target values. For three subgroups—including HS, AIAN and TP—the NAEP-reported estimates include outlier values that are not well approximated by FH estimates (see Figure 4.2.1).

Summary Remarks on Descriptive Statistics of FH Estimates

Despite the greater variance in NAEP estimates for most subgroups, the variances of the FH estimates are typically more similar to the variances of the NAEP-reported estimates than the variances of the estimates produced by the *MICE* technique. For example, the range of NAEP-

reported estimates for the API value is 56 points and 54 points for FH-produced estimates, but only 31 points for *MICE*-produced estimates. Although, the distributions of FH estimates, in general, are relatively similar to NAEP estimates by subgroup, the distribution of FH estimates for the AIAN subgroup in particular is problematic. The range of NAEP estimates for the AIAN subgroup is 34 points while the range of FH estimates for this subgroup is just 13 points.

Table 4.2.2: Descriptive statistics of NAEP-reported vs. FH-produced estimates of mean math achievement by subgroup of interest

	Mean		**SD**		**Min**		**Median**		**Max**		**Range**	
	NAEP	FH	NAEP	FH	NAEP	FH	NAEP	FH	NAEP	FH	NAEP	FH
NHS	265	265	5	3	254	258	266	265	275	272	21	14
HS	268	268	6	5	252	259	268	268	278	277	26	18
SBA	283	283	5	4	270	271	283	283	293	291	23	20
BA	293	293	7	7	276	276	294	294	308	308	32	32
B	259	259	5	4	245	252	258	259	269	267	24	15
H	270	270	4	3	260	264	270	270	279	277	19	13
API	305	305	12	11	275	275	305	307	331	329	56	54
AIAN	259	260	8	3	240	255	260	260	274	268	34	13
TP	282	282	7	5	266	272	282	282	293	291	27	19
EL	246	246	10	8	226	231	246	246	266	263	40	32

Note: Values are rounded to their nearest integer.

Accuracy Statistics for the FH Technique

The overall weighted Mean Absolute Error (wMAE) across subgroups of interest for the FH technique is .49, a dramatic improvement over the MICE technique (1.30). The FH technique's wMAE statistic is smallest (most accurate) for students whose parents' graduated from college (BA; wMAE = 0.20) and largest (least accurate) for Hispanic students (H; wMAE = .91). It should be noted, however, that the FH prediction of mean math achievement for Hispanic students is not inaccurate—at least not in relative terms. Consider, for instance, that the FH technique's wMAE statistic for Hispanic students (0.91) is similar to the most accurate (lowest)

wMAE statistic from the MICE technique (0.85, for the HS subgroup). Table 4.2.3 presents the accuracy statistics for the FH technique by subgroup, as well as across subgroups.

Table 4.2.3: Accuracy statistics for the FH technique by subgroup

Subgroups	weighted Mean Absolute Error (wMAE)	Coverage
Did not finish high school; NHS (n = 48)	.47	.98
Graduated high school; HS (n = 48)	.37	.98
Some education after high school; SBA (n = 48)	.39	1.00
Graduated from college; BA (n = 48)	.20	1.00
Black; B (n = 39)	.76	.97
Hispanic; H (n = 47)	.91	.94
Asian/Pacific Islander; API (n = 30)	.29	1.00
American Indian/Alaskan Native; AIAN (n = 13)	.58	.85
Two or more races; TP (n = 24)	.67	.96
English learner; EL (n = 31)	.34	.97
Overall (n = 376)	.49	.97

The overall coverage statistic across subgroups for the FH technique is .97. Put differently, about 97 percent of the FH-produced estimates of mean math achievement fall within their corresponding target intervals. By contrast, the coverage statistic from the *MICE* technique was 9 percent lower (.88). The FH coverage statistics are perfect for three of ten subgroups (SBA, BA and API), meaning the FH-produced estimates of mean math achievement for these subgroups all fall within their corresponding target intervals. The lowest coverage statistic, by about 9 percentage points, is for the subgroup representing American Indian and Alaskan Native students (.85). For visual representations of the FH-produced estimates of mean math achievement that "hit" and "miss" their corresponding target intervals by subgroup, see Appendix C.

Estimates of Mean Math Achievement with the FLEX CS Technique

The estimates of mean math achievement produced in the FLEX CS technique are formed from combinations of subestimates that vary by subgroup. The estimates for each

subgroup are formed from two or three subestimates (see Table 4.3.1). The Fay-Herriot (FH) and Nearest Neighbor Imputation (NNI) subestimates are used for estimation with all subgroups, while the Multivariate Imputations by Chained Equations (MICE) and Weighted Poststratified Estimator (WPE) subestimates each contribute to FLEX CS estimates for three separate subgroups. The MICE subestimates are used for just three subgroups in the FLEX CS approach, when predictor variables in MICE equations, have a Pearson correlation with the response variable (i.e., observed values from subgroup of interest) of at least .80. The WPE subestimates only factor into the FLEX CS estimates of select states for three subgroups as the source of data used for creating WPE subestimates, the Stanford Educational Data Archive (SEDA; Reardon et al., 2017), includes disaggregated achievement data for only three subgroups of interest in this study. In addition, SEDA only reports achievement data for between thirty to thirty-three states for these three groups (see Table 4.3).

Table 4.3.1: Subestimates used in calculation of FLEX CS estimates by subgroup

Subgroup	Subestimates
NHS	FH, NNI
HS	MICE, FH, NNI
SBA	MICE, FH, NNI
BA	MICE, FH, NNI
B	FH, WPE, NNI
H	FH, WPE, NNI
API	FH, WPE, NNI
AIAN	FH, NNI
TP	FH, NNI
EL	FH, NNI

Note: In the "Subestimates" column: *FH* is a Fay-Heriot subestimate, *NNI* is a Nearest-Neighbor Imputation subestimate, *MICE* is a Multivariate Imputation by Chained Equations subestimate, and *WPE* is Weighed Poststratified Estimator subestimate.

Description of FLEX CS Estimates by Subgroup

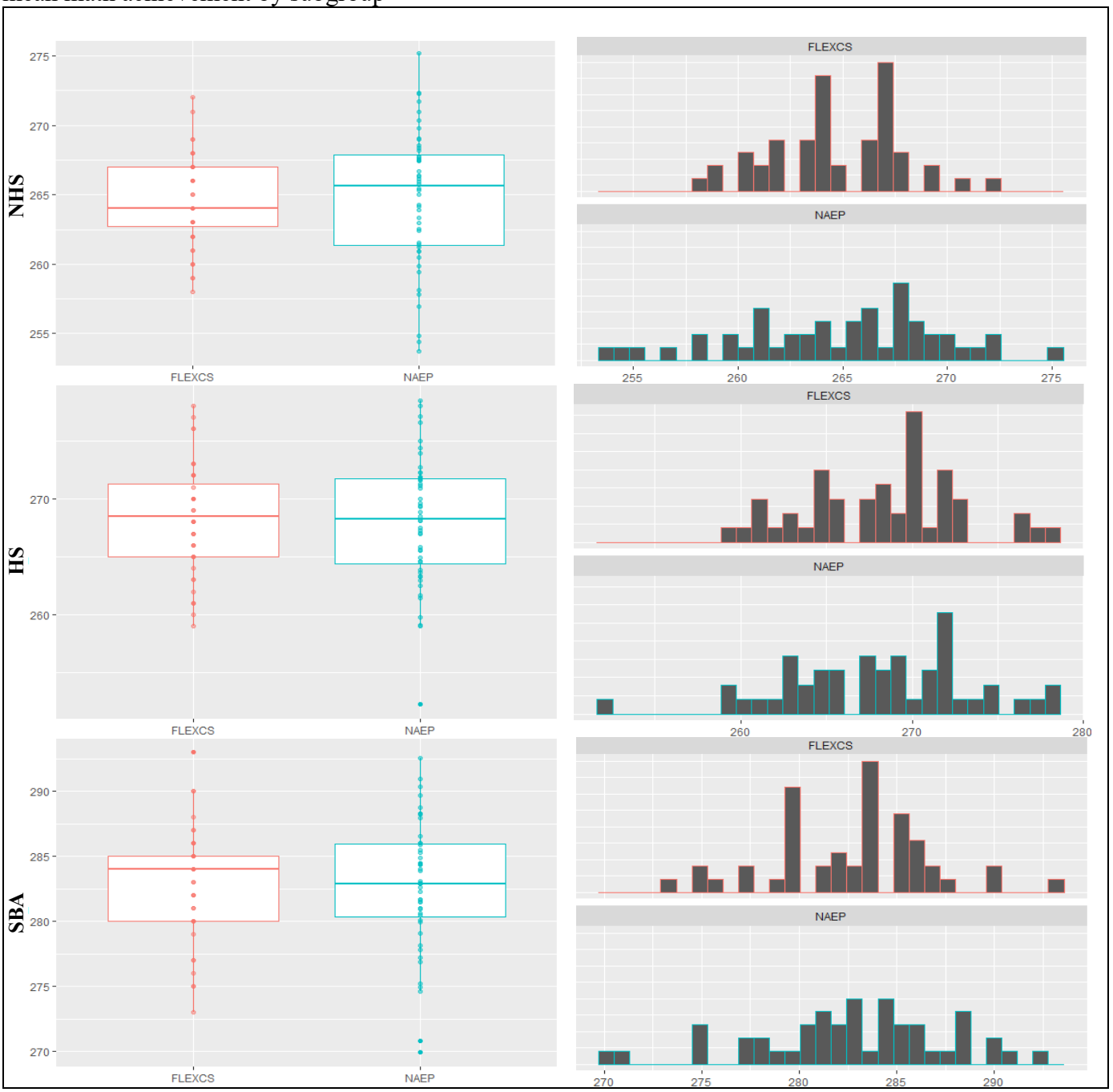
Compared to the estimates produced through the first two techniques, the mean and median values of FLEX CS-produced estimates of mean math achievement are *less* similar to NAEP-reported estimates across subgroups. The mean values of NAEP and FLEX CS estimates

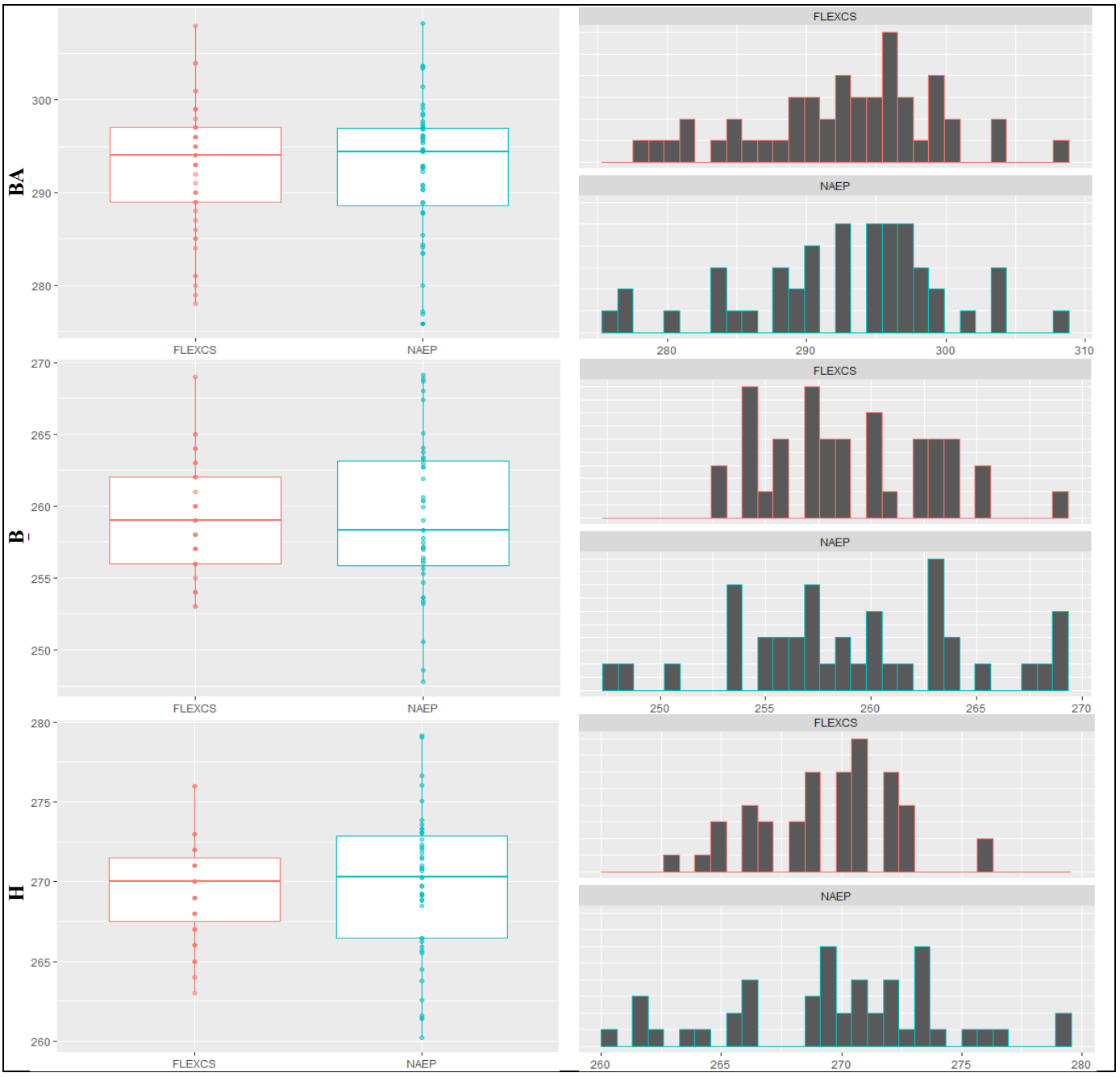
are equal for 7 of 10 subgroups. For the API, AIAN and EL subgroups, the mean FLEX CS estimates are one point greater than NAEP estimates. By contrast, the mean values of NAEP and *MICE* estimates were equal for 9 of 10 subgroups and the mean values of NAEP and *FH* estimates were equal for all subgroups.

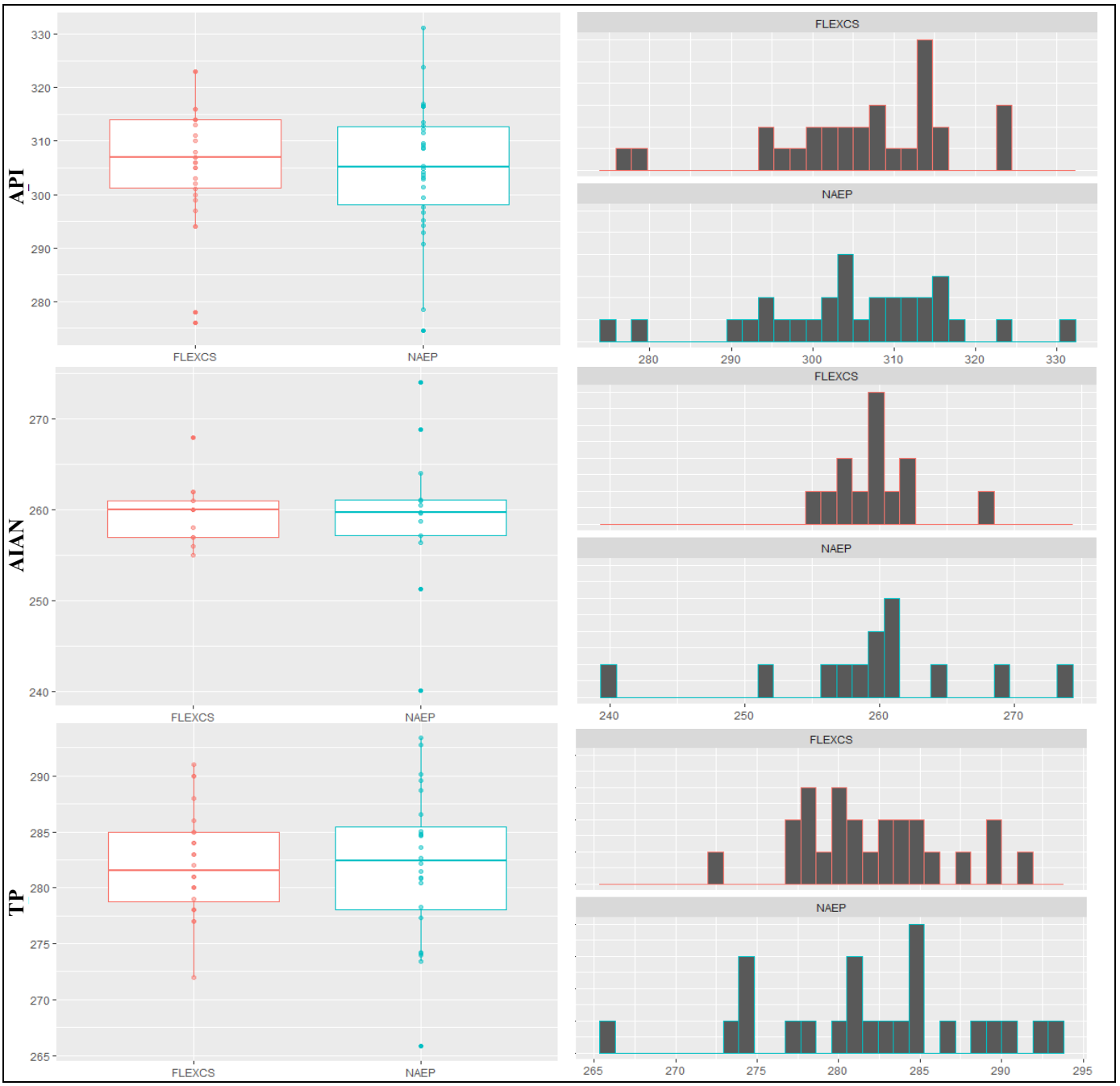
The median values of NAEP and FLEX CS estimates are equal for just 4 of 10 subgroups. For the NHS and API subgroups, the median values are two points apart. For the other 4 subgroups (HS, SBA, B, and EL), the median values of NAEP and FLEX CS estimates are one point apart. For the previous two techniques, median estimate values were equal to the median of NAEP estimate values for 7 of 10 subgroups.

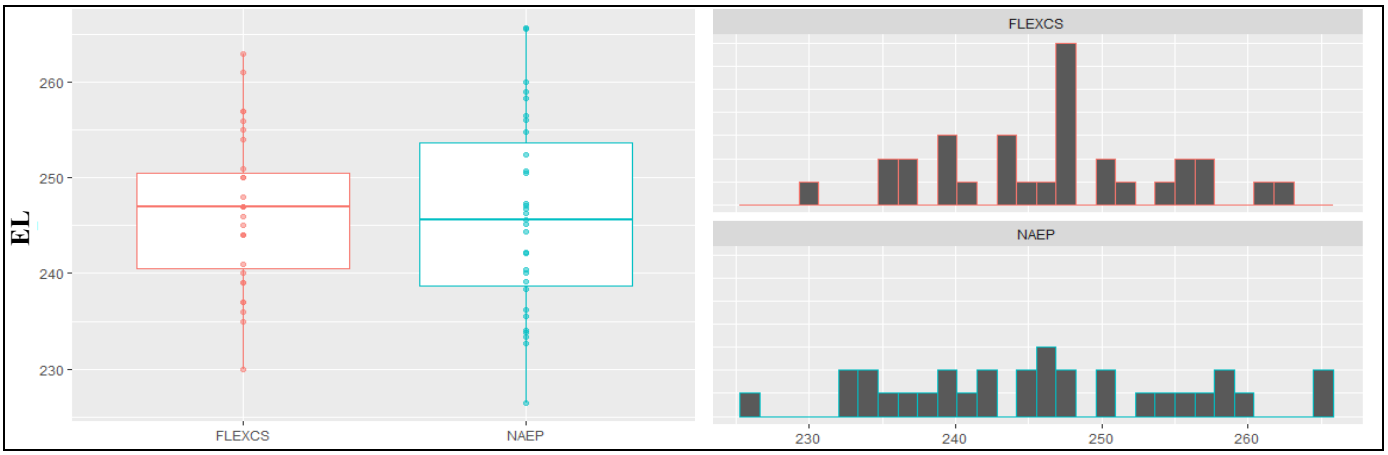
As for the estimates produced through the FH technique, the NAEP-reported estimates of mean math achievement are more dispersed than FLEX CS-produced estimates for all subgroups *except* the subgroup representing students whose parents graduated from college (BA). The standard deviation of estimates produced through the FLEX CS technique is equal to the standard deviation of NAEP-reported estimates for this subgroup (SD = 7, for both NAEP and FLEX CS estimates). Relative to the variance of estimates of mean math achievement produced by the MICE technique, the variance of FLEX CS estimates are generally closer to the variance of NAEP estimates across subgroups. The variances of estimates produced by the FH and FLEX CS techniques are very similar. For each subgroup, as demonstrated in Tables 4.2.2 and 4.3.2, the standard deviation of FH and FLEX CS estimate are equal, rounded to the nearest integer.

Figure 4.3.1: Boxplots and histograms of NAEP-reported vs FLEX CS-produced estimates of mean math achievement by subgroup




BA
B
H
FLEXCS
NAEP
300
290
280
270
265
260
255
250
275
280
310


API
AIAN
TP
FLEXCS
NAEP
330
320
310
300
290
280
270
260
250
240
285
275
265
295



As with the FH technique, the NAEP-reported estimates of mean math achievement for the HS, AIAN and TP subgroups include outliers that are not well approximated by estimates produced by the FLEX CS technique. In addition, the FLEX CS technique does not produce estimates that come particularly close to outlying NAEP-reported estimates for the SBA subgroup (see boxplot for the SBA subgroup in Figure 4.3.1).

Summary Remarks on Descriptive Statistics of FLEX CS Estimates

The mean and median values of FLEX CS-produced estimates of mean math achievement are *less* similar to NAEP-reported estimates across subgroups, compared to mean and median values of MICE- and FH-produced estimates. However, compared to MICE-produced estimates, the range in values of FLEX CS-produced estimates are generally more similar to the range in values of NAEP-reported estimates across subgroups. On the other hand, the range in values of FH and FLEX CS estimates are generally similar across subgroups. One exception is the difference in the range of estimates produced for the Asian Pacific Islander (API) subgroup. The ranges of estimates for the FH and FLEX CS techniques equal 54 and 47, respectively, for the API subgroup, while the range of NAEP estimates for this subgroup is 56.

Table 4.3.2: Descriptive statistics of NAEP-reported vs. FLEX CS-produced estimates of mean math achievement by subgroup of interest

	Mean		**SD**		**Min**		**Median**		**Max**		**Range**	
	NAEP	FLEX CS	NAEP	FLEX CS	NAEP	FLEX CS	NAEP	FLEX CS	NAEP	FLEX CS	NAEP	FLEX CS
NHS	265	265	5	3	254	258	266	264	275	272	21	14
HS	268	268	6	5	252	259	268	269	278	278	26	19
SBA	283	283	5	4	270	273	283	284	293	293	23	20
BA	293	293	7	7	276	278	294	294	308	308	32	30
B	259	259	5	4	245	253	258	259	269	269	24	16
H	270	270	4	3	260	263	270	270	279	276	19	13
API	305	306	12	11	275	276	305	307	331	323	56	47
AIAN	259	260	8	3	240	255	260	260	274	268	34	13
TP	282	282	7	5	266	272	282	282	293	291	27	19
EL	246	246	10	8	226	230	246	247	266	263	40	33

Note: Values are rounded to their nearest integer.

Accuracy Statistics for the FLEX CS Technique

The overall weighted Mean Absolute Error (wMAE) across subgroups of interest for the FLEX CS technique is .70, which represents considerably greater accuracy than the MICE technique's wMAE (1.30), but lower than the FH technique's wMAE (.49). The FLEX CS technique's wMAE statistic is smallest (most accurate) for students who are English learners (.45) and greatest for Black students (.98). On the other hand, while the overall wMAE statistic for the *FH* technique is smaller (more accurate) than the wMAE for the FLEX CS technique (.49 vs .70), the FLEX CS estimates were more accurate for Hispanic students (H subgroup) and nearly identical to the FH technique's wMAE statistic for students identifying with two or more races (TP subgroup). The FLEX CS technique's wMAE statistic for the H subgroup equals .85, while the FH wMAE statistic equals .91. The FLEX CS technique's wMAE statistic for the TP subgroup equals .67, while the FH wMAE statistic for TP students equals .68. Table 4.3.3 presents the accuracy statistics for the FLEX CS technique by subgroup, as well as across subgroups.

Table 4.3.3: Accuracy statistics for the FLEX CS technique by subgroup

Subgroup	weighted Mean Absolute Error (wMAE)	Coverage
Did not finish high school; NHS (n = 48)	.61	.96
Graduated high school; HS (n = 48)	.69	.98
Some education after high school; SBA (n = 48)	.74	.96
Graduated from college; BA (n = 48)	.64	1.00
Black; B (n = 39)	.98	.90
Hispanic; H (n = 47)	.85	.98
Asian/Pacific Islander; API (n = 30)	.69	.90
American Indian/Alaskan Native; AIAN (n = 13)	.62	.85
Two or more races; TP (n = 24)	.68	.96
English learner; EL (n = 31)	.45	.90
Overall (n = 376)	.70	.95

The overall coverage statistic across subgroups of interest for the FLEX CS technique is .95. This means that about ninety-five percent of the FLEX CS-produced estimates of mean math achievement fall within their corresponding target intervals. In comparison, the coverage statistic from the MICE technique was seven percent lower (.88) and the coverage statistic from the FH technique was two percent higher (.97). FLEX CS coverage by subgroup is perfect (1.00) for just one subgroup, for students whose parents graduated from college (BA). By contrast, the MICE and FH techniques had coverage statistics of 1.00 for two and three subgroups, respectively. It should be noted that all three techniques have coverage statistics of 1.00 for the BA subgroup. The FLEX CS coverage statistic is smallest for the AIAN subgroup (.85). The AIAN subgroup was also associated with the smallest coverage statistic for the FH technique. Illustrations of FLEX CS-produced estimates of mean math achievement that "hit" and "miss" their corresponding target intervals by subgroup are provided in Appendix C.

Research Question Analyses

This section revisits the research questions put forth at the outset of analyses, the rationale behind each question, and the criteria proposed to answer them. Three separate but

related research questions shaped the analyses. The first is a general question that asks, in absolute terms, whether any of the techniques under evaluation could defensibly be used by NAEP researchers to estimate the mean math achievement of subgroups that are unreported by the NAEP program. Specifically, *is it reasonable to use any of the techniques examined in this study, based on benchmarks established through a simulation analysis, to estimate subgroup math achievement on State NAEP when sample sizes do not permit direct estimation*? The second asks how accurately, in relative terms, the techniques predict mean achievement. Specifically, *how do the techniques compare with respect to maximizing accuracy, according to accuracy measures used in this study (weighted Mean Absolute Error and coverage)*? The third asks about the predictive accuracy of techniques by subgroup. Specifically*, how do the techniques vary in their ability to predict achievement per subgroup*?

Research Question 1 - Is it reasonable, based on benchmarks established through a simulation analysis, to use any of the techniques examined in this study to estimate subgroup math achievement on State NAEP when sample sizes do not permit direct estimation?

The central question of this study is whether one or more of the techniques can reasonably be applied to the estimation of mean math achievement of subgroups on State NAEP when direct sampling does not yield samples of at least 62 students. This question is addressed through an analysis of coverage statistics. A technique is considered reasonable to use if, *across* subgroups of interest, at least 95 percent of the technique's predicted estimates of mean math achievement fall within corresponding target intervals, and *per* subgroup of interest, at least 80 percent fall within corresponding target intervals. These rates, considered markers of successful

prediction, were established based on results from a simulation analysis of example NAEP data from the *EdSurvey* package (Bailey et al., 2019) in R.[44]

The answer to this central question is yes: two of the three evaluated techniques could reasonably be applied to the estimation of mean subgroup achievement on State NAEP when direct sampling does not yield at least 62 students identifying with the subgroup (according to the criteria established in this study). While the coverage statistics from the MICE technique do *not* meet the criteria that would render the technique reasonable to use, the coverage statistics from the FH and FLEX CS techniques do meet the criteria. The overall coverage statistics from the FH and FLEX CS techniques equal .97 and .95, respectively, both of which are greater than or equal to the benchmark value of 0.95. Meanwhile, the smallest coverage statistic values by subgroup from the FH and FLEX CS techniques are both .85 (both for the AIAN subgroup), which is greater than the benchmark value of .80.

It should be carefully noted, however, that the FH and FLEX CS techniques must be applied to NAEP achievement data from other years and for the NAEP *Reading* assessment before definitive recommendations can be made for their use in practice. That is, judgements regarding their general promise and utility cannot be cast without first evaluating the predictive accuracy of these technique with achievement data from different years and NAEP Reading.

Research Question 2 - How do the techniques compare with respect to maximizing accuracy, according to accuracy measures used in this study (weighted Mean Absolute Error and coverage)?

Addressing the second research question permits a conclusion to be drawn regarding which techniques perform best in *relative* terms. In other words, it asks which technique performs *best* in terms of its ability to accurately predict mean subgroup achievement. The

[44] The simulation analysis and results are described at length in Chapter 3. The statistical code can be found on the author's GitHub page.

answer to this question has heightened importance since it was determined that two techniques (i.e., FH & FLEX CS) meet the criteria to be considered adequate for use. This question is answered through an analysis of weighted Mean Absolute Error (wMAE) statistics—a weighted measure of mean absolute differences between predicted values (i.e., technique-produced values) and target values (i.e., NAEP-reported values).

The simple and straightforward answer is that the FH technique is more accurate than the FLEX CS technique. The overall wMAE statistics (across subgroups) from the FH and FLEX CS techniques are .49 and .70, respectively. For the FH technique, measures of wMAE range from .20 to .91 across subgroups, while FLEX CS measures of wMAE range from .45 to .98. The more nuanced answer to this second research question is that the FH techniques *tends* to be more accurate than the FLEX CS technique—a point treated in greater detail by the answer to the third research question.

Research Question 3 - How do the techniques vary in their ability to predict achievement per subgroup?

The third research question prompts an examination into whether prediction accuracy for each technique varies as a function of the subgroup. The importance of this question is twofold. First, the answer helps to ascertain whether a technique that is generally successful in predicting mean subgroup achievement overall can also accurately predict mean achievement for each subgroup. A technique that cannot successfully predict mean achievement for each subgroup does not have the same practical appeal to practitioners and researchers, nor does it have firm standing as a defensible prediction technique. It raises the suspicion that the technique's predictive success occurs by chance, at least to some extent. Second, the answer to this research question offers an opportunity to learn about the relative usefulness of the different inputs (e.g., predictor variables) used to estimate mean math achievement from the different techniques.

For all but one subgroup, the FH technique most accurately predicts mean math achievement. In more specific terms, the wMAE statistics associated with the FH technique are smallest for nine subgroups. On the other hand, the wMAE statistic associated with the FLEX CS technique is smallest for the subgroup representing Hispanic students (H). This finding, however, does not mean that the FH technique inadequately predicts the mean math achievement of Hispanic students. The coverage statistic associated with the FH technique for Hispanic students is .94, which is well above the benchmark of .80 deemed to represent an acceptable coverage rate for any given subgroup. This last finding does however suggest that the WPE subestimate for Hispanic students constructed from SEDA (Reardon et al., 2017), which forms part of the FLEX CS mean estimates for Hispanic students, contributes to accurate prediction, while the WPE subestimates for other subgroups are less helpful. In addition, the greater general accuracy of the FH technique over FLEX CS suggests that the NNI subestimates, which were used in the latter technique, are not particularly helpful.

Applying the FH technique to Unreported Achievement Data

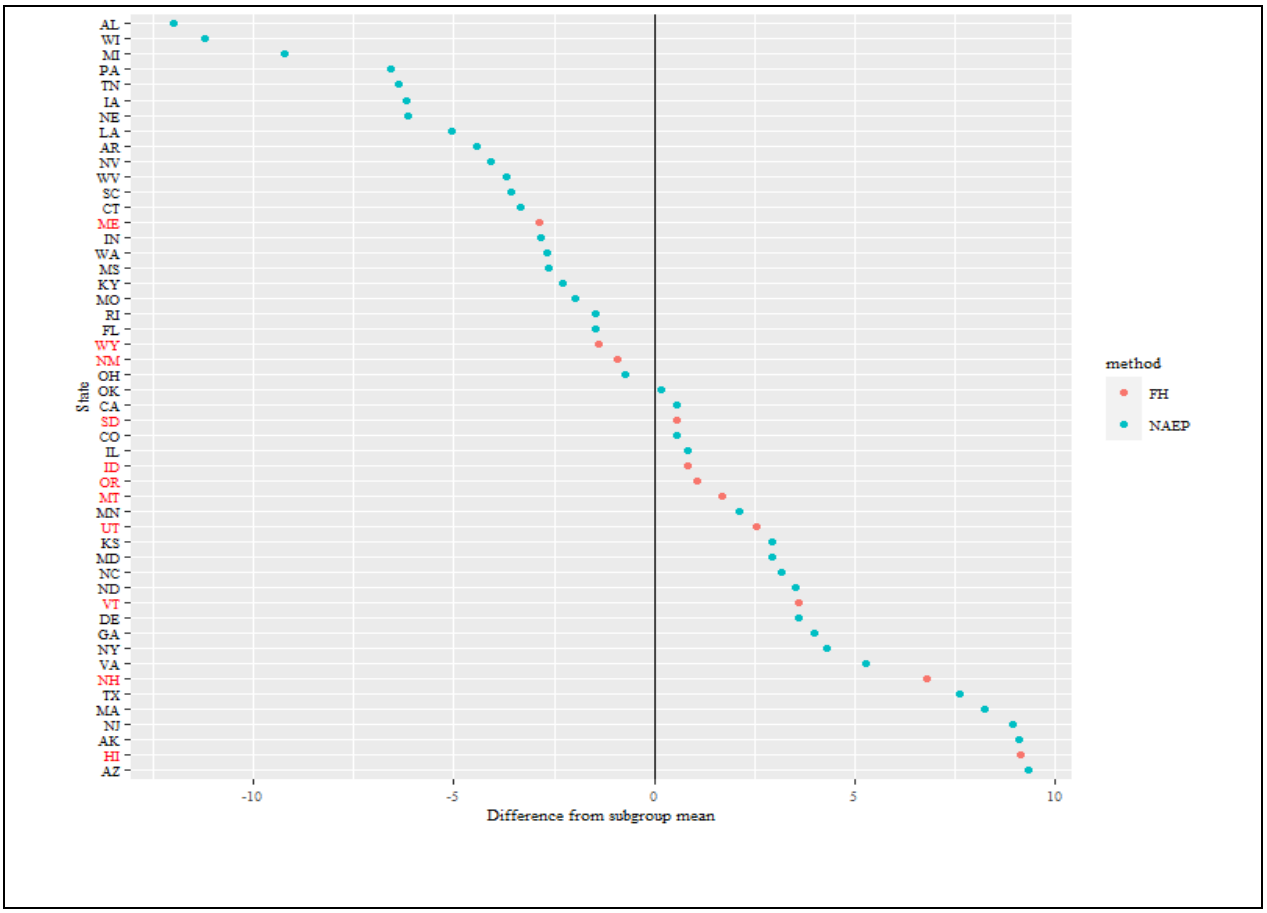
To examine whether using the FH technique in practice yields reasonable results, the technique was applied to the subgroup representing Black students (B). This subgroup was chosen for a number of reasons. The legacy of slavery and discrimination against Black people in American history has long hampered the educational opportunities and advancement of Black students, and that has led to special interest in supporting Black students' learning and achievement. In addition, NAEP does not report an estimate of mean math achievement for this subgroup in 11 of 50 states from the test sample,[45] providing an opportunity to evaluate whether

[45] The 11 states are: Hawaii (HI), Idaho (ID), Maine (ME), Montana (MT), New Hampshire (NH), New Mexico (NM), Oregon (OR), South Dakota (SD), Utah (UT), Vermont (VT), and Wyoming (WY).

the FH technique produces reasonable estimates of mean math achievement for 11 states instead of just two or three.

The mean estimates (EBLUPs) for the 11 states calculated with the FH technique appear reasonable. For example, the range of these 11 imputed values is contained within the range of the 39 observed values (i.e., NAEP-reported estimates). Further, the average of the 11 FH-produced mean estimates and 39 NAEP-reported mean estimates are similar—262 and 259, respectively. Figure 4.4 is a scatter plot that displays the differences between each state's estimate of mean math achievement of Black students and the average mean math achievement estimate of Black students across all 50 states. The 11 imputed FH estimates, represented by red dots, are located within the distribution of NAEP estimates, represented by blue dots, and they do not cluster in any particular region of the overall distribution of estimates.

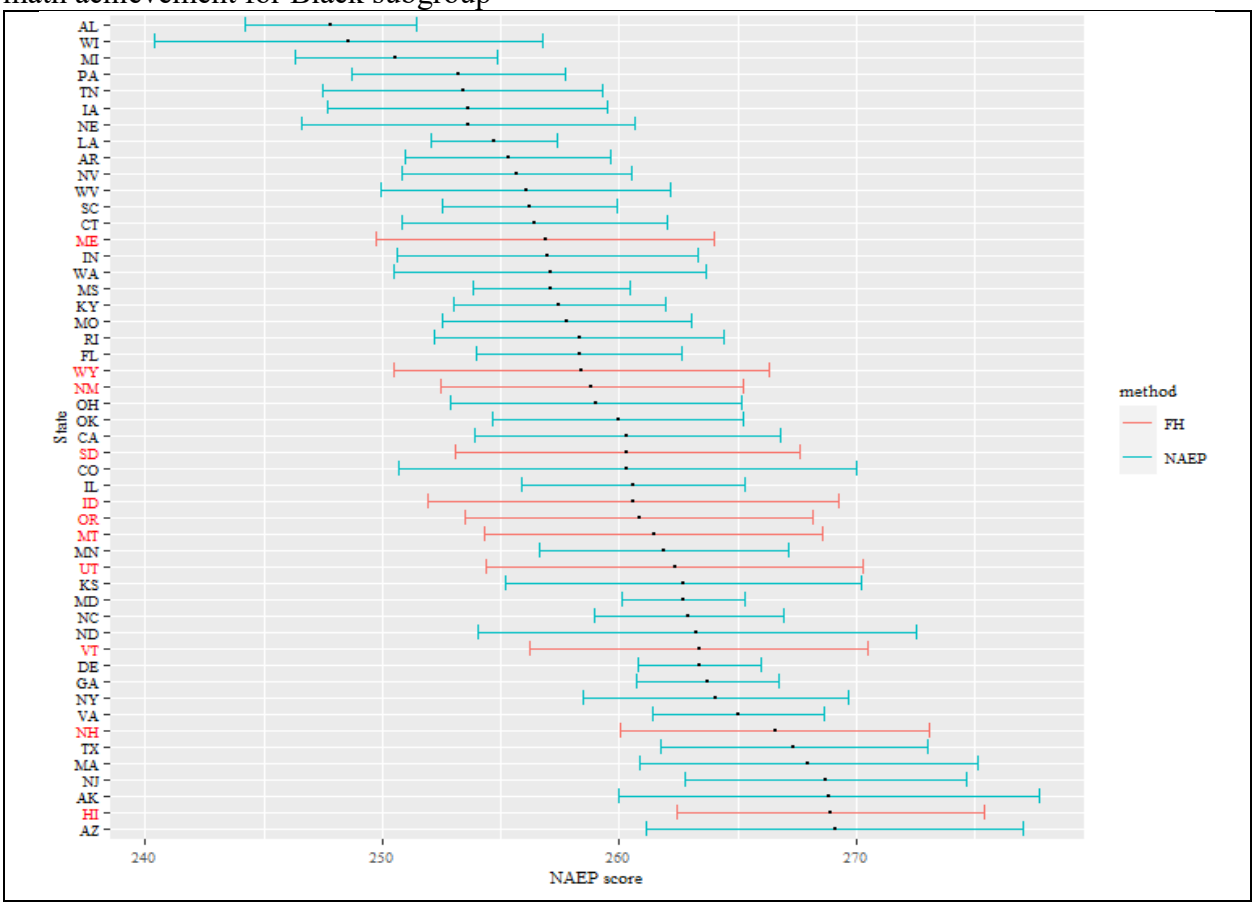
Figure 4.4. Scatter plot of the differences between each state's estimate of mean math achievement of Black students and the nationwide average of estimates of mean math achievement of Black students



The values of mean variance estimates for the 11 states calculated with the FH technique are also acceptable. The standard errors associated with the 11 EBLUPs range in value from 3.3 to 4.4, with a mean of 3.7 and median of 3.6. By contrast, the standard errors associated with the 39 NAEP-reported estimates of mean math achievement range from 1.3 to 4.2 with a mean of 2.6 and median of 2.5.[46] Figure 4.5 displays the 95-percent confidence intervals of mean estimates for all 50 states, with intervals representing FH estimates color-coded red.

[46] Mean estimates and standard errors of all 50 states for the B subgroup are provided in Table B.11 in Appendix B.

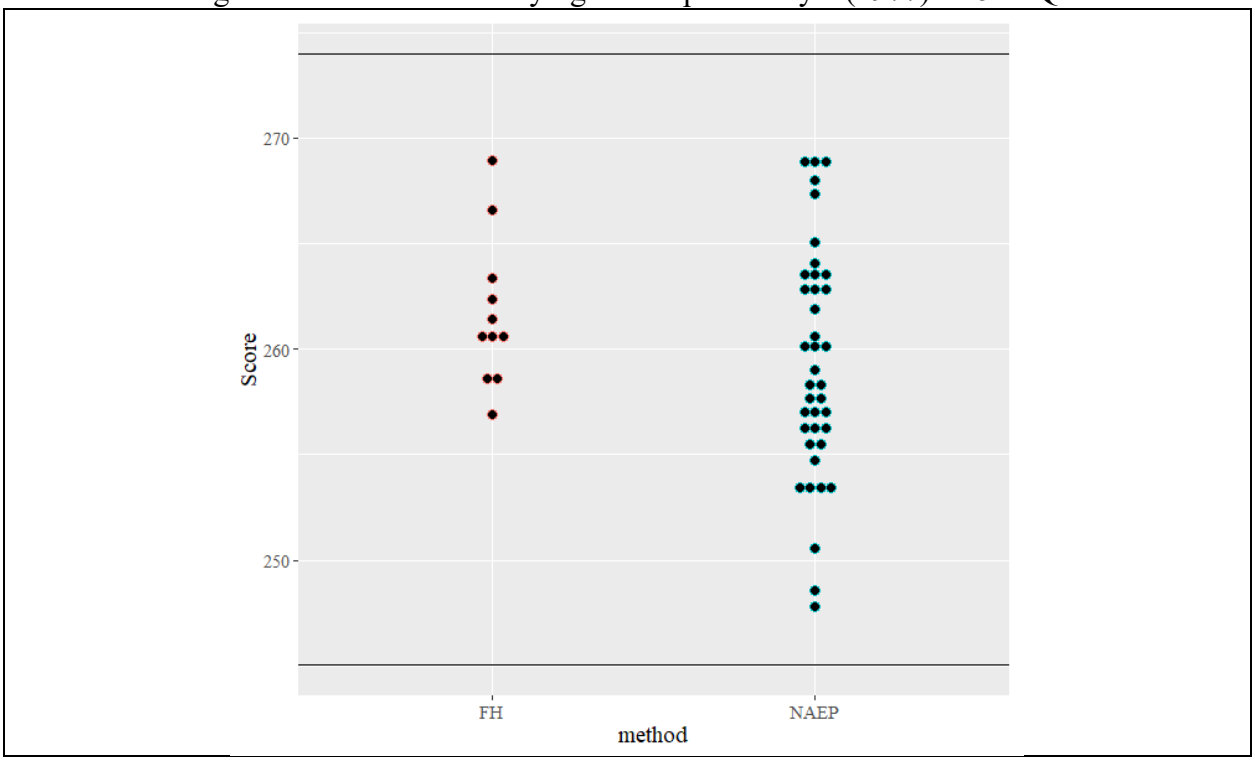
Figure 4.5: 95-percent confidence intervals of FH (red) and NAEP (blue) estimates of mean math achievement for Black subgroup



Although it is clear that FH estimates are generally calculated less precisely, in comparison to NAEP-reported estimates, the size of the 11 standard errors are within the range of standard error values that the NAEP program is accustomed to reporting. Consider, for instance, that the maximum values of the 11 FH and 39 NAEP-reported standard errors are similar, 4.4 and 4.2, respectively. In practice, we can expect some FH-produced estimates, relative to design-based estimates, to be calculated imprecisely. It is unavoidable that, in some cases, only a few students will be sampled and available for the computation of the direct-estimate component of the EBLUP.

As an additional step to evaluate the credibility of the 11 imputed estimates of mean achievement produced through the FH technique, a test was undertaken to examine whether any of the 11 imputed values would be considered outlying values according to Tukey's (1977) "$1.5 \times IQR$" rule. Figure 4.6 presents juxtaposed dot plots of FH and NAEP-reported estimates of mean math achievement. The region bounded by the horizontal lines represents a range of non-outlying values, according to the"$1.5 \times IQR$" rule, based on the interquartile range of NAEP-reported estimates. Results indicate that the values of FH estimates are non-outlying, which serves as additional evidence that the imputed estimates of mean math achievement calculated through the FH technique are plausible.

Figure 4.6. Dotplot of estimates of mean achievement by estimation method, with horizontal lines demarcating boundaries of non-outlying values per Tukey's (1977) "1.5 x IQR" rule.

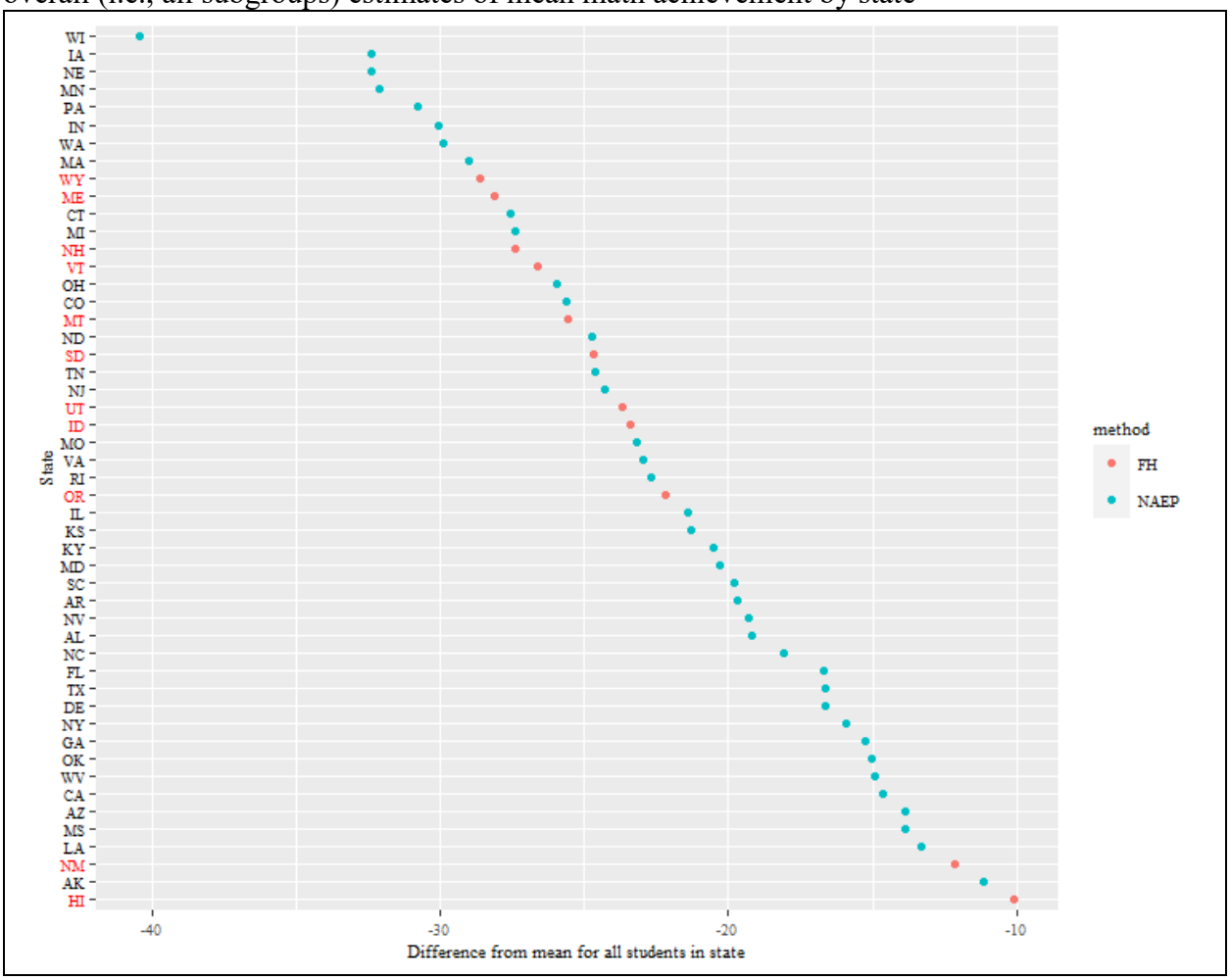


Note: The black horizontal lines demarcate the lower and upper bounds, respectively 245 and 274, of the reference range used for assessing the plausibility of predicted values for the B subgroup.

An additional step to ensure the 11 FH estimates represent credible values of mean math achievement for the B subgroup, *overall* mean math achievement estimates were compared to mean math achievement estimates of Black students for all 50 states. Figure 4.7 illustrates the gaps (i.e., differences) in mean achievement estimates between *all* students and Black students by state. The differences calculated by subtracting estimates of mean math achievement of Black students produced through FH from the estimates of mean math achievement of all students, as reported by NAEP, are color-coded red. These results lend additional credibility to the 11 estimates calculated through the FH technique since the observed gaps between the achievement estimates of Black students, calculated through FH, and estimates of all students would be expected to be similar, in direction and magnitude, to the gaps observed between Black students and all students in the 39 remaining states.[47]

[47] Although Figures 4.4 and 4.7 appear similar, they present different comparisons. Figure 4.4 displays the differences between the mean achievement estimates of Black students in individual states and the mean achievement of Black students nationwide (i.e., the vertical line). Figure 4.7 displays differences between mean achievement estimates of Black students and all students (i.e., all subgroups) within each state.

Figure 4.7: Differences between estimates of mean math achievement of Black students and overall (i.e., all subgroups) estimates of mean math achievement by state



Chapter 5: Discussion

The purpose of this study was to ascertain whether one or more of three techniques were appropriate choices for estimating subgroup performance on State NAEP. Whether a technique was considered appropriate was determined by a set of rules that reflected how close predicted values produced by the techniques were to values reported by the NAEP program. The three techniques were progressively more complex with respect to the data entered and the manner in which estimates were constructed. The first technique was as an adaptation of a form of multiple imputation, Multivariate Imputation by Chained Equations (MICE). The second was a Small Area Estimation technique, the Fay-Herriot model (FH). The third was a form of cross-survey analysis, a technique referred to in this study as Flexible Cross-Survey Analysis (FLEX CS). This chapter revisits the findings, discusses limitations, and offers recommendations for future research.

Review of the Findings

A technique is considered reasonably accurate in predicting mean subgroup achievement and potentially suitable for actual implementation if *both* its aggregate coverage statistic is 0.95 or higher and by-subgroup coverage statistics are 0.80 or higher. Two of the three techniques evaluated in this study, FH and FLEX CS, appear suitable for use in practice. The MICE technique does not.

The aggregate coverage statistic associated with MICE is 0.88, which is lower than the aggregate benchmark value of 0.95. In addition, the coverage statistics associated with MICE for

four separate subgroups (i.e., API, AIAN, TP, and EL) are less than the benchmark value of 0.80 set for individual subgroups.

Table 5.1: Subgroup and aggregate measures of *wMAE* and *coverage* by technique

	MICE	FH	FLEX CS
Did not finish high school (n = 48)			
Weighted Mean Absolute Error (wMAE)	1.01	.47	.61
Coverage	.92	.98	.96
Graduated high school (n = 48)			
Weighted Mean Absolute Error (wMAE)	0.85	.37	.69
Coverage	1.00	.98	.98
Some education after high school (n = 48)			
Weighted Mean Absolute Error (wMAE)	1.05	.39	.74
Coverage	.98	1.00	.96
Graduated college (n = 48)			
Weighted Mean Absolute Error (wMAE)	1.16	.20	.64
Coverage	1.00	1.00	1.00
Black (n = 39)			
Weighted Mean Absolute Error (wMAE)	1.35	.76	.98
Coverage	.90	.97	.90
Hispanic (n = 47)			
Weighted Mean Absolute Error (wMAE)	1.23	.91	.85
Coverage	.98	.94	.98
Asian/Pacific Islander (n = 30)			
Weighted Mean Absolute Error (wMAE)	2.73	.29	.69
Coverage	.57	1.00	.90
American Indian/Alaskan Native (n = 13)			
Weighted Mean Absolute Error (wMAE)	1.41	.58	.62
Coverage	.69	.85	.85
Two or more races (n = 24)			
Weighted Mean Absolute Error (wMAE)	1.09	.67	.68
Coverage	.67	.96	.96
English learner (n = 31)			
Weighted Mean Absolute Error (wMAE)	1.85	.34	.45
Coverage	.65	.97	.90
Total (n = 376)			
Weighted Mean Absolute Error (wMAE)	1.30	.49	.70
Coverage	.88	.97	.95

Comparing the FH and FLEX CS techniques, the former appears more promising. In this study, aggregate measures of weighted Mean Absolute Error (wMAE) and coverage both indicate that the FH estimates of mean math achievement are more accurate than the FLEX CS estimates. Although the aggregate coverage statistic is only marginally greater for the FH technique (0.97 vs. 0.95), the aggregate wMAE statistic indicates the FH technique is substantially more accurate. The aggregate wMAE statistics for the FH and FLEX CS techniques

equal 0.49 and 0.70, respectively. That is, on average, the weighted difference between FH estimates of mean math achievement and NAEP-reported estimates across all subgroups equals 0.49 points, while the average weighted difference between FLEX CS and NAEP estimates equals 0.70, corresponding to a 43 percent increase in the wMAE measure.[48] On the NAEP scale, the overall MAE (i.e., unweighted) statistics for FH and FLEX CS equal about 1.5 points and 2.1 points, respectively. For reference, overall estimates of mean math achievement across states in the test sample have a standard deviation of about 6.7 points and a range of 30 points, from 267, in Alabama, to 297, in Massachusetts (U.S. Department of Education, 2021).

By subgroup, the FH technique also generally outperforms FLEX CS. The wMAE and coverage statistics associated with the FLEX CS technique indicate greater accuracy for only one of ten subgroups—the subgroup representing Hispanic students. Still, the accuracy statistics for the Hispanic subgroup are marginally more favorable for the FLEX CS technique compared to FH. The wMAE statistics for the FH and FLEX CS techniques equal 0.91 and 0.85, while their coverage statistics equal 0.94 and 0.98, respectively.

Since the FH technique is the best performing technique, it is proposed that follow-up research on the estimation of mean subgroup achievement on NAEP be focused on the implementation of the FH technique. Further, the analysis discussed at the end of the previous chapter makes clear that the FH technique can yield reasonable results when applied to actual missing achievement data. When applied to the Black subgroup, the FH technique produced estimates of mean math achievement for the 11 states unreported by NAEP that are comparable

[48] The weighting step in the calculation of wMAE produces aggregate differences that are smaller in magnitude than actual average differences between predicted and target values on the NAEP scale. This occurs because absolute differences between predicted and target values are divided by the standard error associated with the target value (i.e., the NAEP-reported estimate).

to estimates that would have been reasonably expected given the NAEP program had sampled at least 62 Black students from the 11 states.

Evaluation of the MICE Technique

The MICE technique adequately predicts the mean math achievement of the parental level of education subgroups. However, the technique inadequately predicts mean math achievement estimates of race and ethnicity subgroups. The coverage statistics from MICE for this second set of subgroups range from .57 to .98, though these ranges belie the fact that, for four of these five subgroups, the coverage statistic is equal to a value below the .80 benchmark. The MICE technique also fails to adequately predict the mean math achievement estimates for the English learners subgroup (.65).

Overall, the MICE technique meets neither the coverage statistic benchmark of .95 across subgroups, nor the benchmark of .80 per subgroup. It should be noted, however, that the MICE technique's general ability to accurately predict mean achievement was likely inadvertently hamstrung by the decision to use the Predictive Mean Matching (PMM) approach instead of normal linear regression for some subgroups. The PMM approach was used for estimation with the SBA, BA and AIAN subgroups and exacerbated the MICE technique's tendency to produce estimates that are biased toward the state averages of mean achievement across subgroups.

Regardless of PMM's role in the prediction of mean achievement, the MICE technique would have proved inadequate. Consider, for instance, that the MICE technique is especially inaccurate in its prediction of mean achievement for the API subgroup, for which PMM was not used. In summary, the MICE technique *cannot be recommended* for use in practice.

Evaluation of the FH Technique

The FH technique, like the MICE technique, adequately predicts the mean math achievement of parental level of education subgroups. The coverage statistics from the FH technique for these subgroups equal .98 for the NHS and HS subgroups, and 1.00 for the SBA and BA subgroups. This means FH-produced estimates of mean math achievement only miss their associated target intervals on two occasions across the four parental level of education subgroups. For the NHS subgroup, the FH-produced estimate for Connecticut is about 1 point greater than the upper bound of its corresponding target interval (see Table C.2.1, Appendix C). For the HS subgroup, the FH-produced estimate for Alabama is slightly greater than the upper bound of its corresponding target interval (see Table C.2.2, Appendix C).

The FH technique also adequately predicts the mean math achievement of race and ethnicity subgroups, although not as well as it predicts the parental level of education subgroups. The coverage statistics from FH for this second set of subgroups range from .85, for the AIAN subgroup, to 1.00, for the API subgroup. For the English learners subgroup (EL), the FH coverage statistic equals .97. Notwithstanding the less than desirable prediction accuracy for the AIAN subgroup, the FH technique *could be recommended* for use in practice.

It should be noted, however, that while the coverage statistic of .85 is comfortably greater than the benchmark value of .80 per subgroup for a technique to be considered acceptable, this coverage statistic of .85 is notably lower than the coverage statistics for the remaining race and ethnicity subgroups, which otherwise range from .94 to 1.00. The FH-produced estimates for the AIAN subgroup miss their target intervals for Utah and Wisconsin, the lowest- and highest-achieving states according to NAEP's estimates. The NAEP-reported estimates of mean achievement in Utah and Wisconsin are 240 and 274, while the corresponding FH estimates

equal 256 and 262, respectively. It should be noted, however, that the NAEP-reported mean achievement values for Utah and Wisconsin are estimated imprecisely relative to other states. The standard errors reported by NAEP for Utah and Wisconsin are 9.0 and 7.0 points, which represent the largest standard errors reported by NAEP for the AIAN subgroup. It is conceivable that the true parameter values for these states are closer to their FH estimates. There is substantial overlap in the confidence intervals of the NAEP and FH estimates for these two states. For Utah, the 95-percent confidence intervals range from 222 to 258 and 247 to 265. For Wisconsin, these intervals range from 260 to 288 and 254 to 270.

Evaluation of the FLEX CS Technique

The FLEX CS technique, like the MICE and FH techniques, adequately predicts mean math achievement for the parental level of education subgroups. The coverage statistics from FLEX CS for these subgroups equal .96 for the NHS and SBA subgroups, .98 for the NHS subgroup, and 1.00 for the BA subgroup. While these coverage rates are relatively high, and well above the benchmark of .80, they are not quite as high as the corresponding rates from the FH technique. This small difference between FH and FLEX CS coverage statistics for parental level of education subgroups indicates that the MICE and NNI subestimates that factor into FLEX CS estimates for these subgroups, in addition to FH subestimates, do not provide improved accuracy. Analysis of wMAE statistics from FH and FLEX CS techniques for the parental level of education subgroups corroborate this last point. The FH technique's wMAE statistics, compared to the FLEX CS technique's, are notably smaller for these four subgroups.

The FLEX CS technique, like the FH technique, also satisfactorily predicts mean math achievement for the race and ethnicity subgroups. The coverage statistics from FLEX CS for the race and ethnicity subgroups range from .85, for the AIAN subgroup, to .98, for the H subgroup.

For the English learners subgroup (EL), the FH coverage statistic equals .96. For one subgroup, representing Hispanic students, FLEX CS estimates are more accurate than FH estimates. FLEX CS estimates, however, are not more accurate for the other race and ethnicity subgroups. The FLEX CS technique, while not generally as accurate as the FH technique, *could also be recommend* for use in practice.

Two additional points about FLEX CS results bear mentioning. First, like the FH technique, the subgroup for which FLEX CS is least effective in predicting mean math achievement is the one representing American Indian and Alaskan Native students (AIAN). This finding could signal that estimates for this particular subgroup are generally the most difficult to predict. A logical explanation is that the samples used for computing direct estimates for this subgroup are much smaller, relative to the samples available for the other subgroups. In this study, the sample size, rounded to the nearest 10, used for computing direct estimates for the AIAN subgroup equal 10. The sample sizes used for the other subgroups range from 30 to 50.

Second, with the exception of one subgroup (H), combining estimates produced with the FH technique with estimates produced with the other estimation methods (i.e., MICE, WPE, NNI), does not appear to be a useful strategy. It is worth unpacking the limited utility of WPE and NNI subestimates, in particular, since the use of these types of estimates is unique to estimation with FLEX CS.

Regarding WPE subestimates, it is helpful to remember that these subestimates are used for only three subgroups of interest (i.e., B, H & API) and subsets of states because of limited availability of achievement data in the Stanford Educational Data Archive (SEDA). Nonetheless, a number of interesting findings regarding WPE subestimates emerge from scrutinizing FLEX CS estimates across these three subgroups. First, the improvement (i.e., decrease) in wMAE for

the FLEX CS technique, relative to FH, for the Hispanic subgroup is largely driven by estimates for California and Florida. Interestingly, these states have relatively large Hispanic populations. The target value (i.e., NAEP-reported estimate) for Hispanic students in California equals 263, and the corresponding FH and FLEX CS estimates equal 271 and 265, respectively. The target value for Hispanic students in Florida equals 272, and the corresponding FH and FLEX CS estimates equal 267 and 271, respectively. The coverage statistic for the Hispanic subgroup is greater for the FLEX CS technique, compared to FH, because the FLEX CS estimate for Hispanic students in California is located within its target interval, but the FH estimate for Hispanic students in California is not (as illustrated in Figures C.2.6 and C.3.6 in Appendix C).

The only other instance in which a FLEX CS estimate, constructed in part with a WPE subestimate, is substantially more accurate than an FH estimate is for the Black subgroup in Georgia. The target value for Black students in Georgia equals 264, and its corresponding FH and FLEX CS estimates equal 258 and 262, respectively. Interestingly, Georgia is among the states with the largest proportions of Blacks students. This finding, paired with the finding regarding the greater accuracy of FLEX CS estimates for Hispanic students in California and Florida, provides some indication, albeit limited, that using SEDA data could be helpful for estimation in states where minority subgroups represent a relatively large proportion of the general state population.

Regarding estimation with Nearest Neighbor Imputation (NNI), it should similarly be noted that the potential for NNI subestimates to contribute to FLEX CS estimates of mean math achievement was limited in this study. First, there are only 12 states (i.e., 6 pairs) considered similar enough to warrant NNI subestimation in the calculation of their FLEX CS estimates. These are the pairs of states described as *sibling* states. Second, for some subgroups of interest,

certain estimates of sibling states cannot be used as NNI subestimates because they are unreported in NAEP publications, meaning they cannot be used since they do not exist. To understand this point, consider that certain pairs of sibling states are, in relative terms, demographically homogenous (e.g., Iowa-North Dakota, Kansas-Nebraska) and that the NAEP program would have experienced difficulty in sufficiently sampling students from certain subgroups (e.g., API, AIAN, EL) in these pairs of states.

In addition to limits on their applicability, the NNI subestimates are relatively inaccurate. As can be deduced by examining tables in Appendix B, NNI subestimates are seldom more accurate (i.e., closer to the target value) than corresponding MICE and FH estimates. For two pairs of sibling states, NAEP-reported estimates of mean math achievement are especially dissimilar across subgroups. NAEP estimates of mean math achievement for Oklahoma and New Jersey respectively tend to be considerably greater than estimates for Alabama and Connecticut.

Limitations of the Study

Several study limitations deserve mention and are discussed below.

Indirect Comparisons of Techniques

In a sense, the predictive performance of the three techniques evaluated in this study are not *directly* compared since they do not draw on the same sources of data for prediction. It can instead be argued that three separate analytic *approaches* are directly compared. This perspective asserts that separate techniques are not compared since they use different data for prediction and these data represent a potential confounding factor in the evaluation of each technique's relative predictive accuracy. It is conceivable that the relative accuracy of these techniques is not so

much influenced by the algorithms or processes governing the techniques but the data that they incorporate for prediction. Results from this study could have been different with different data.

Estimating the Mean Achievement of Intersections of Subgroups

This study does not attempt to estimate the mean achievement of intersections of NAEP subgroups (e.g., Black male students, Asian students with limited English proficiency). While estimating the mean achievement of intersections of subgroups is beyond the scope of this current study, estimating the mean achievement of more narrowly defined groups of students is a worthy endeavor. Many students belong to multiple historically underserved and underperforming groups of students (e.g., American Indian students living in poverty, racial minorities with limited English proficiency) and the social factors that place these students at an academic disadvantage are multifaceted.

Consider, for instance, that recent reform efforts in the United States have been designed to specifically support the academic success of Black *male*s, who, in the aggregate, experience far less favorable outcomes than other groups of students—including higher dropout, suspension and expulsion rates (Chen, 2020; Lynch, 2017). It is crucially important that researchers and policymakers have a better understanding of the learning and achievement of particular intersections of subgroups, especially those that are chronically underserved and underperforming. While NAEP's public-use Data Explorer computes mean achievement estimates of intersections of subgroups across states, the availability of these estimates is greatly limited by NAEP's minimum sample size policy (i.e., rule-of-62) for reporting.

Predicting Mean Achievement with Proxy Data with the FH Technique

Most of the administrative data used to construct predictor variables for the regression models in the FH approach are *indirect* measures (i.e., proxy variables) of the factors they are

intended to represent. Ideally, the predictor variables would specifically represent state-level characteristics of grade 8 students in 2015 in public schools, the target population for the desired inference. However, the administrative data available for constructing these predictor variables are imperfect matches, primarily because the available data are not disaggregated by grade level.

Limited SEDA Data

Using SEDA data (Reardon et al., 2017), which provides NAEP-referenced mean scale scores for subgroups by district to estimate mean subgroup achievement at the state-level, would seem to provide a promising solution to the research problem that this dissertation addresses. Unfortunately, these district-level data are only available through SEDA for three of ten subgroups of interest in this study (B, H & API). Further, the district-level data for these three subgroups are only available for a limited number of the 50 states. For the test sample (i.e., mean math achievement of 8th graders in 2015), district-level estimates from SEDA are available for 34 of the 50 states. In practice, where the research aim would be to estimate the mean achievement of state-subgroup pairs that *are not reported* by NAEP, using SEDA data might have even more limited use, as the states for which these district-level estimates are missing in SEDA also tend to be the states for which mean subgroup achievement are not reported by NAEP.

Recommendations for Future Research

An obvious next step is to evaluate the predictive accuracy of the Fay-Herriot (FH) technique with math achievement data from other years and a parallel set of analyses for the NAEP Reading data, using the same criteria from this study to determine whether the accuracy observed with other test samples are reasonably successful. Further, improving the prediction of mean achievement when sampling does not permit direct estimation should be treated an as

ongoing effort. The prediction techniques explored and data used in this study are not exhaustive. Different techniques and sources of data can be identified that may improve prediction. An intriguing method to explore is the Spatio-Temporal Fay-Herriot (STFH) model, an extension of the common Fay-Herriot model, which also makes use of area-level data from different time periods for estimation (Molina & Marhuenda, 2015). In the context of this research, applying the STFH model would involve borrowing mean achievement data from preceding and subsequent NAEP testing years, as available, to strengthen prediction.

The set of estimates in greatest need of improvement in this study are those for the American Indian or Alaskan Native (AIAN) subgroup. Efforts to improve the accuracy of these estimates should involve evaluating estimates of mean achievement produced from different combinations of predictor variables for this subgroup. In this study, since there are only 13 target values (i.e., NAEP-reported estimates) for this subgroup, just two predictor variables are used to calculate regression-estimates, *FER* and *%BA*, representing the economic circumstances of students' families and parental level of education, respectively. Considering that aggregate measures of these factors (e.g., state-level variables) tend to be highly correlated, it is conceivable that prediction for this subgroup could be improved by combining FER and %BA into a composite variable, which then offers an opportunity for another variable to enter the model without the risk of overfitting. A promising predictor, for instance, would be one that measures states' proportions of AIAN students who attend Bureau of Indian Education (BIE) schools or schools with relatively high concentrations of AIAN students, "high density public schools" in NAEP terminology.[49] Previous research (Milne, 2016; Ninneman, Deaton & Francis-

[49] The NAEP program defines *high density public schools* as schools with AIAN enrollment of at least 25% percent. Note, these do not include schools run by the Bureau Indian Education (BIE).

Begay, 2017) demonstrates that AIAN students in BIE and high-density schools tend to be lower achieving.

For all subgroups, it is worth examining the changes in the accuracy of FH estimates after substituting predictors used in this study for the FH approach with predictors used in the regression models from the MICE approach. More specifically, it may prove helpful to use NAEP-reported estimates of mean achievement from one subgroup to produce the regression estimators of another subgroup through the FH technique. For instance, it could be helpful to replace one of the predictor variables currently used to predict the mean achievement of a parental level of education subgroup with a variable representing NAEP-reported estimates of mean achievement from a separate parental level of education subgroup.

In instances in which the mean achievement estimate of just one subgroup from a category of subgroups is missing (unreported), it is reasonable to estimate the missing mean achievement value through algebraic steps, given the number of students in each subgroup, as a proportion of students from all subgroups that form the demographic category, are known. Data that permit a researcher to reasonably approximate the distribution of students across a category of subgroups by year, grade-level and state may be available in databases such as the Common Core of Data.

Finally, it is worth exploring whether applying techniques from the ever-expanding field of machine learning would improve prediction. Exploring the utility of machine learning to address this study's research problem is a logical next step since machine learning techniques are squarely focused on prediction problems. Among the multitude of machine learning algorithms that exist, the Random Forests algorithm should be among the first to be evaluated since it is one of the most widely used and popular algorithms, and represents an extension of regression

trees—which are commonly used for predicting continuous outcome measures (Hastie, Tibshirani & Friedman, 2009; Irizarry, 2020).

Final Conclusions

Results indicate that two of the three techniques studied in this dissertation would be suitable for use in practice and that the Fay-Herriot (FH) technique is particularly appealing. The practical significance of these findings is that they provide initial evidence to support the idea that it could be defensible and appropriate to use a prediction model to calculate estimates of mean achievement when the NAEP program is unable to sample enough students from a particular state and subgroup to calculate direct estimates. The benefit of implementing such a policy is that it would provide a more complete understanding of how states support the learning and achievement of different subgroups of students, including underserved and underperforming subgroups of students, whose mean achievement is frequently unreported because these subgroups often represent small proportions of states' general populations.

Results from estimating the mean math achievement of Black students in states unreported by NAEP through the FH technique serve as an example of the potential benefits of using a prediction model when direct estimation is impermissible. The results, for instance, provide some evidence to suggest that the achievement gap between Black students and all students by state is smallest in Hawaii, a state for which NAEP did not report the mean math achievement of 8th grade Black students in 2015. This kind of finding is significant as it has the potential to inspire follow-up research regarding the reasons a relatively small achievement gap is observed in Hawaii. Although the political appetite may never exist to publish model-based estimates of mean achievement alongside direct estimates in traditional NAEP reports, precedence exists to provide *full* reporting of state subgroup achievement through supplemental

materials, much in the same way “Full Population Estimates” have been published since 2005 (U.S. Department of Education, 2020b).

References

Altman, D. G. (1991). Practical Statistics for Medical Research. London, UK: Chapman & Hall.

Anderson, M. (2015, April). Chapter 1: Statistical Portrait of the U.S. Black Immigrant Population. Pew Research Center. Retrieved from https://www.pewsocialtrends.org/2015/04/09/chapter-1-statistical-portrait-of-the-u-s-black-immigrant-population/

Austin, P. C. & Steyerberg, E. W. (2015). The number of subjects per variable required in linear regression analyses. Journal of Clinical Epidemiology, 68(6), 627-636.

Azur, M., Stuart, E., Frangakis, C., & Leaf, P. (2011). Multiple Imputation by Chained Equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20(1), 40-49.

Bailey, P., C'deBaca, R., Emad, A., Huo, H., Lee, M., Liao, Y., Lishinski, A., Nguyen, T., Xie, Q., Yu, J., & Zhang, T. (2019). EdSurvey: Analysis of NCES Education Survey and Assessment Data. R package version 2.3.2. https://CRAN.Rproject.org/package=EdSurvey

Bains, N. (2009). Standardization of Rates. Association of Public Health Epidemiologists in Ontario (APHEO). Retrieved from http://core.apheo.ca/resources/indicators/Standardization%20report_NamBains_FINALMarch16.pdf

Barnes, S., Lindborg, S., & Seaman, J. (2006). Multiple imputation techniques in small sample clinical trials. *Statistics in Medicine*, 25, 233-245.

Bengtsson, H. (2018). matrixStats: Functions that Apply to Rows and Colums of Matrices (and to Vectors). R package version 0.54.0. https://CRAN.R-project.org/package=matrixStats

Beresovsky, V. & Hsiao, J. (2014). Methodological aspects of small area estimation from the National Electronic Health Records Survey (NEHRS). National Center for Education Statistics. Retrieved from https://nces.ed.gov/FCSM/pdf/A2_Beresovsky_2013FCSM.pdf

Berliner, D. (2006). Our impoverished view of education reform. *Teachers College Record*, 108(6).

Berliner, D. (2013). Effects of Inequality and Poverty vs. Teachers and Schooling on America's Youth. *Teachers College Record*, 115(12).

Best, N., Richardson, S., Clarke, P., & Gomez-Rubio, V. (2008). A Comparison of Model-based Methods for Small Area Estimation. National Centre for Research Methods. Retrieved from https://www.researchgate.net/publication/268059411_A_Comparison_of_model-based_methods_for_Small_Area_Estimation

Bohrnstedt, G., Kitmitto, S., Ogut, B., Sherman, D., and Chan, D. (2015). School Composition and the Black–White Achievement Gap (NCES 2015-018). U.S. Department of Education, Washington, DC: National Center for Education Statistics. Retrieved [date] from http://nces.ed.gov/pubsearch.

Bourque, M. L. (2004). A History of the National Assessment Governing Board. Pp. 201-231 in Jones, L.V. & Olkin, I. (Eds.). (2004). The Nation's Report Card: Evolution and Perspectives. Bloomington, IN: Phi Delta Kappa Educational Foundation.

Braun, H. & Jones, D. (1984). Use Empirical Bayes Methods in the Study of the Validity of Academic Predictors of Graduate School Performance. ETS Research Report Series, 2, i-83.

Braun, H. & Kirsch, I. (2016). Introduction: Opportunity in America—Setting the Stage. In Kirsh, I. & Braun, H. (eds.), *The Dynamics of Opportunity in America*. Princeton, NJ: Educational Testing Service.

Braun, H. (2016). The Dynamics of Opportunity in America: A Working Framework. In Kirsh, I. & Braun, H. (eds.), *The Dynamics of Opportunity in America*. Princeton, NJ: Educational Testing Service.

Carter, D. J. (2008). On spotlighting and ignoring racial group members in the classroom. In Mica Pollock (Ed). Everyday anti-racism: Getting real about race in the classroom. New York: The New Press, pp. 230-234.

Chen, F. (2020, July). 'My Brother's Keeper' Seeks to Give African-American Boys a Boost. Public School Review. Retrieved from https://www.publicschoolreview.com/blog/my-brothers-keeper-seeks-to-give-african-american-boys-a-boost

Chenevert, R., Gottschalck, A., Klee, M., & Zhang, X. (2017). Where the Wealth Is: The Geographic Distribution of Wealth in the United States. U.S. Census Bureau: Social, Economic and Housing Statistics Division. Retrieved from https://www.census.gov/content/dam/Census/library/working-papers/2017/demo/FY2016-129.pdf

Chingos, M. (2015). *Breaking the curve: Promises and pitfalls of using NAEP data to assess the state role in student achievement.* Urban Institute. Retrieved from

http://gateway.proquest.com/openurl?url_ver=Z39.88-2004&res_dat=xri:policyfile&rft_dat=xri:policyfile:article:00181689

Chromy, J. R., Finker, A. L., & Horvitz, D. G. (2004). Survey Design Issues. Pp. 383-425 in Jones, L.V. & Olkin, I. (Eds.). (2004). The Nation's Report Card: Evolution and Perspectives. Bloomington, IN: Phi Delta Kappa Educational Foundation.

Cleger-Tamayo, S., Fernandez-Luna, J. M., & Huete, J. F. (2012). On the use of Weighted Mean Absolute Error in Recommender Systems. Workshop on Recommendation Utility Evaluation: Beyond RMSE (RUE 2012), held in conjunction with ACM RecSys 2012. September 9, 2012, Dublin, Ireland. Paper retrieved from http://ceur-ws.org/Vol-910/paper5.pdf

Code of Federal Regulations, Title 34—Education, Part 99. Family Educational and Privacy Rights, (34CFR99). Washington, DC: GPO Access e-CFR. Retrieved from https://www.ecfr.gov/cgi-bin/text-idx?tpl=/ecfrbrowse/Title34/34cfr99_main_02.tpl

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences. New York, NY: Routledge Academic

Collins, L., Schafer, J., & Chi-Ming, K. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. Psychological Methods, 6(4), 330-351.

Czajka, J. (2016). Small Area Estimates Produced by the U.S. Federal Government: Methods and Issues. Paper presented at the "Small Area Estimation Conference," August 17-19, 2016, Maastricht, the Netherlands. Retrieved from https://www.mathematica-mpr.com/our-

publications-and-findings/publications/small-area-estimates-produced-by-the-us-federal-government-methods-and-issues

Dahl, G. B. & Lochner, L. (2012). The Impact of Family Income on Child Achievement: Evidence from the Earned Income Tax Credit. *American Economic Review*, American Economic Association, 102(5), 1927-56

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, Series B, 39(1), 1-38. Retrieved from https://www.jstor.org/stable/2984875?seq=1#page_scan_tab_contents

Drakos, G. (2018, August). How to Select The Right Evaluation Metric for Machine Learning Models: Part 1 Regression Metrics. Retrieved from https://towardsdatascience.com/how-to-select-the-right-evaluation-metric-for-machine-learning-models-part-1-regrression-metrics-3606e25beae0

Education Week Research Center. (2015). Quality Counts 2015: State Report Cards Map. Retrieved from https://www.edweek.org/ew/qc/2015/2015-state-report-cards-map.html?intc=EW-QC15-LFTNAV

Elliot, E. & Phillips, G. (2004). A View from NCES. Pp. 233-249 in Jones, L.V. & Olkin, I. (Eds.). (2004). The Nation's Report Card: Evolution and Perspectives. Bloomington, IN: Phi Delta Kappa Educational Foundation.

Eskelson, B., Temesgen, H., Lemay, V., Barrett, T., Crookston, T., & Hudak, A. (2009). The roles of nearest neighbor methods in imputing missing data in forest inventory and monitoring databases. USDA Forest Service. University Of Nebraska-Lincoln Faculty

Publications. Retrieved from https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1216&context=usdafsfacpub

Foley, D. (2004). Ogbu's theory of academic disengagement: its evolution and its critics. *Intercultural Education*, 15(4), 385-397.

Galwey, N. K. (2014). Introduction to Mixed Modelling: Beyond Regression and Analysis of Variance. John Wiley & Sons: Chichester, UK.

Gay, G. (2002). Culturally responsive teaching in special education for ethnically diverse students: Setting the stage. International Journal of Qualitative Studies in Education, 15(6), 613-629.

Ghosh, M. & Rao, J.N.K. (1994). Small Area Estimation: An Appraisal. *Statistical Science*, 9(1), 55-76.

Glass G. V., McGaw B., Smith M. L. (1981). Meta-Analysis in Social Research. Beverly Hills, CA: Sage

Gomes, J. H. F., Paiva, A. P., Costa, S. C., Balestrassi, P.P., & Paiva, E. J. (2013). Weighted Multivariate Mean Square Error for processes optimization: A case study on flux-cored arc welding for stainless steel claddings. *European Journal of Operational Research*, 226 (3), 522-535.

Gomez-Rubio, V., Best, N., Richardson, S., Clarke, P., & Li, G. (2010). Bayesian Statistics for Small Area Estimation. Retrieved from https://www.semanticscholar.org/paper/Bayesian-Statistics-Small-Area-Estimation-G%C3%B3mez-Rubio-Best/6fcdc61e715bf81c1849a5d0b523d026d41f4894

Graham, J. W. & Schafer, J.L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In Statisitical Strategies for Small Sample Research, ed. R Hoyle, 1:1-29. Thousand Oaks, CA: Sage.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory. *Prevention Science*, 8, 206-213.

Graham, J. W. (2009). Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology*, 60, 549-576.

Hair, Joseph F., Jr; Black, William C.; Babin, Barry J.; and Anderson, Rolph E. (2009) *Multivariate Data Analysis* (7th Ed.). Upper Saddle River, NJ: Prentice Hall.

Hanushek, E., Peterson, P., & Woesmman, L. (2013). Endangering Prosperity: A Global View of the American School. Washington, DC: Brookings Press.

Hedges, L. V. & Bandeira de Mello, V. (2013). *NAEP Validity Studies: A Validity Study of the NAEP Full Population Estimates*. Washington, DC: American Institutes for Research. https://www.air.org/sites/default/files/downloads/report/A_Validity_Study_of_Full_Population_Estimates_NAEP_0.pdf

Hedges L. V., Olkin I. (1985). Statistical methods for meta-analysis. San Diego, CA: Academic Press.

Hastie T., Tibshirani R., Friedman J. (2009). The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York, NY.

Irizarry, R. (2020). Introduction to Data Science: Data Analysis and Prediction Algorithms with R. Retrieved from https://rafalab.github.io/dsbook/

Johnson, D. R. & Young, R. (2011). Toward Best Practices in Analyzing Datasets with Missing Data: Comparisons and Recommendations. *Journal of Marriage and Family*, 73, 926-945.

Kreuter, F., Eckman, S. Maaz, K., & Watermann, R. (2010). Children's Reports of Parents' Education Level: Does it Matter Whom You Ask and What You Ask About? *Survey Research Methods*, 4(3), 127-138.

Ladson-Billings, G. (2006). From the achievement gap to the education debt: Understanding achievement in U.S. schools. Address to American Education Research Association.

Lapoint, A. (2004). A New Design for a New Era. Pp. 185-199 in Jones, L.V. & Olkin, I. (Eds.). (2004). The Nation's Report Card: Evolution and Perspectives. Bloomington, IN: Phi Delta Kappa Educational Foundation.

Lipsey, M. W. (2001). Practical meta-analysis. London, UK: SAGE

Little, R. (1988) Missing-Data Adjustments in Large Surveys, *Journal of Business & Economic Statistics*, 6(3), 287-296.

Longford, N.T. (2005). Missing Data and Small-Area Estimation: Modern Analytical Equipment for the Survey Statistician. New York, NY: Springer.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9(1), 1-19.

Lynch, M. (2017, April). A Guide to Ending the Crisis Among Young Black Males. The Edvocate. Retrieved from https://www.theedadvocate.org/guide-ending-crisis-among-young-black-males/

Magadin de Kramer, R. (2016). Evaluation of Cross-Survey Research Methods for the Estimation of Low-Incidence Populations. Retrieved from ProQuest LLC. (ProQuest Number: 10247646)

Manna, P. (2013). Centralized governance and student outcomes: Excellence, equity, and academic achievement in the U.S. states. *Policy Studies Journal* 41(4): 682-705.

McKown, C. & Weinstein, R. (2008). Teacher expectations, classroom context, and the achievement gap. Journal of School Psychology, 46(3), 235-261. Retrieved from https://www.sciencedirect.com/science/article/pii/S0022440507000416

Milne, D. (2016). Exploring the American Indian/Alaska Native 8th Grade Patterns inMathematics Achievement in Arizona and South Dakota. Dissertation retrieved from https://core.ac.uk/download/pdf/151481749.pdf

Molina, I. & Marhuenda, Y. (2015). sae: an R package for Small Area Estimation. *The R Journal*, 7(1).

Mullis, I. (2004). Assessing Writing and Mathematics. Pp. 361 – 380 in Jones, L.V. & Olkin, I. (Eds.). (2004). The Nation's Report Card: Evolution and Perspectives. Bloomington, IN: Phi Delta Kappa Educational Foundation.

Musu-Gillette, L., Robinson, J., McFarland, J., Kewal-Ramani, A., Zhang, A., and Wilkinson-Flicker, S. (2016). Status and Trends in the Education of Racial and Ethnic Groups 2016 (NCES 2016-007). U.S. Department of Education, National Center for Education Statistics. Washington, DC. Retrieved from https://nces.ed.gov/pubs2016/2016007.pdf

National Research Council. (2000). Small Area Income and Poverty Estimates: Priorities for 2000 and Beyond. Washington, DC: The National Academies Press.

National School Board Association. (2015, May). Learning to Read, Reading to Learn. Learning First Alliance. Retrieved from https://learningfirst.org/learning-read-reading-learn

Ninneman, A.M., Deaton, J., and Francis-Begay, K. (2017). National Indian Education Study 2015 (NCES 2017-161). Institute of Education Sciences, U.S. Department of Education, Washington, DC. Retrieved from https://nces.ed.gov/nationsreportcard/subject/publications/studies/pdf/2017161.pdf

O'Dwyer, L. M., and Parker, C. E. (2014). A primer for analyzing nested data: multilevel modeling in SPSS using an example from a REL study (REL 2015–046). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Northeast & Islands. Retrieved from http://ies.ed.gov/ncee/edlabs.

Ogbu, J. & Simons, H. (1998). Voluntary and involuntary minorities: a cultural-ecological theory of school performance with some implications for education, *Anthropology and Education Quarterly*, 29(2), 155–188.

Ogbu, J. (2003). Black American students in an affluent suburb: A study of academic disengagement. New Jersey: Lawrence Erlbaum.

Olkin, I. (2004). Interviews. Pp. 251-289 in Jones, L.V. & Olkin, I. (Eds.). (2004). The Nation's Report Card: Evolution and Perspectives. Bloomington, IN: Phi Delta Kappa Educational Foundation.

Pfefferman, D. (2002). Small Area Estimation – New Developments and Directions. *International Statistical Review*, 70(1), 125-143.

Ponomarenko, N.N., Krivenko, S. S., Egiazarian, K., & Lukin, V. V. (2010). Weighted mean square error for estimation of visual quality of image denoising methods. Paper presented at Workshop on Video Processing and Quality Metrics for Consumer Electronics. Paper retrieved from

https://www.researchgate.net/publication/273444562_Weighted_mean_square_error_for_estimation_of_visual_quality_of_image_denoising_methods

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Rao, J. N. K. (2012). Small Area Estimation: Methods and Applications. Paper presented at the Seminar "Applications of Small Area Estimation Techniques in the Social Sciences," October 3-5, 2012, Iberoamerican University, Mexico City. Retrieved from https://www.inegi.org.mx/eventos/2012/Ciencias_sociales/doc/Rao_Mexico_City_slides.pdf

Rao, J. N. K. (2013). Small Area Estimation: Methods, Applications and New Developments. Paper presented at the NTTS 2013 Conference, Brussels, March 2013. Retrieved from https://ec.europa.eu/eurostat/cros/system/files/9A01_Keynote_Rao-v2_0.pdf

Rao, J. N. K. & Molina, I. (2015). Small Area Estimation (2nd Edition). Hoboken, NJ: Wiley.

Reardon, S., Ho. A., Shear, B., Fahle, E., Kalogrides, D., & DiSalvo, R. (2017). Stanford Education Data Archive (Version 2.0). http://purl.stanford.edu/db586ns4974.

Reardon, S., Kalogrides, D. & Ho, A. (2019). Validation Methods for Aggregate-Level Test Scale Linking: A Case Study Mapping School District Test Score Distributions to a Common Scale. *Journal of Educational and Behavioral Statistics*, #(#).

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York, NY: John Wiley and Sons.

Scott, L.A., and Ingels, S.J. (2007). Interpreting 12th-Graders' NAEP-Scaled Mathematics Performance Using High School Predictors and Postsecondary Outcomes From the National Education Longitudinal Study of 1988 (NELS:88) (NCES 2007-328). National

Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Schafer, J. L. & Graham, J. W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods*, 7(2), 147-177.

Selden, R. (2004). Making NAEP State-by-State. Pp. 195-199 in Jones, L.V. & Olkin, I. (Eds.). (2004). The Nation's Report Card: Evolution and Perspectives. Bloomington, IN: Phi Delta Kappa Educational Foundation.

Shavelson, R., & Towne, L. (2002). Features of education and education research. Scientific research in education. Washington, DC: National Academy Press.

StataCorp. (2019). *Stata Statistical Software: Release 16*. College Station, TX: StataCorp LLC.

Steele, C. & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African-Americans. *Journal of Personality and Social Psychology, 69*, 797-811.

Steele, C., Spencer, S., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. *Advances in experimental social psychology*, 34, 379-440.

Tabachnick, B. G., & Fidell, L. S. (2007). Using multivariate statistics (5th ed.). Boston, MA: Allyn & Bacon/Pearson Education.

Taylor, R. (1990). Interpretation of the Correlation Coefficient: A Basic Review. *Journal of Diagnostic Medical Sonography*, 6, 35-39.

Tukey, J. W. (1977). Exploratory data analysis. Reading, PA: Addison-Wesley.

U.S. Department of Education. (2001). Education achievement and Black-White Inequality. Washington, DC: Department of Education. Washington, DC: Institute of Education

Sciences, National Center for Education Statistics. Retrieved from https://nces.ed.gov/pubs2001/2001061.PDF

U.S. Department of Education. (2002, November). How States Join. Washington, DC: Institute of Education Sciences, National Center for Education Statistics. Retrieved from https://nces.ed.gov/nationsreportcard/about/statejoin.aspx

U.S. Department of Education. (2008, December). NAEP Technical Documentation: NAEP Data Explorer. Washington, DC: Institute of Education Sciences, National Center for Education Statistics. Retrieved from https://nces.ed.gov/nationsreportcard/tdw/database/data_tool.asp

U.S. Department of Education. (2010). Statistical Methods for Protecting Personally Identifiable Information in Aggregate Reporting. SLDS Technical Brief: Guidance for Statewide Longitudinal Data Systems (SLDS). Washington, DC: Institute of Education Sciences, National Center for Education Statistics. Retrieved from https://nces.ed.gov/pubs2011/2011603.pdf

U.S. Department of Education. (2015a, December). NAEP Frequently Asked Questions (FAQs). Washington, DC: Institute of Education Sciences, National Center for Education Statistics. Retrieved from https://osse.dc.gov/sites/default/files/dc/sites/osse/page_content/attachments/NAEP%20FAQs.pdf

U.S. Department of Education. (2015b, November). NAEP Sample Design, Weights, Variance Estimation, IRT Scaling, and Plausible Values. Washington, DC: Institute of Education

Sciences, National Center for Education Statistics. Retrieved from https://nces.ed.gov/training/datauser/NAEP_04/assets/NAEP_04_Slides.pdf

U.S. Department of Education. (2017, July). Timeline for National Assessment of Educational Progress (NAEP) Assessments from 1969 to 2024. Washington, DC: Institute of Education Sciences, National Center for Education Statistics. Retrieved from https://nces.ed.gov/nationsreportcard/about/assessmentsched.aspx

U.S. Department of Education. (2018, April). NAEP Reporting Groups. Washington, DC: Institute of Education Sciences, National Center for Education Statistics. Retrieved from https://nces.ed.gov/nationsreportcard/reading/interpret_results.aspx#repgroups

U.S. Department of Education. (2020a, September). ELSi: Elementary/Secondary Information System. Washington, DC: Institute of Education Sciences, National Center for Education Statistics. Retrieved from https://nces.ed.gov/ccd/elsi/

U.S. Department of Education. (2020b, February). Full Population Estimates. Washington, DC: Institute of Education Sciences, National Center for Education Statistics. Retrieved from https://nces.ed.gov/nationsreportcard/about/fpe.aspx

U.S. Department of Education. (2020c, December). Statistical Standards Program: Publication of products using restricted-use data. Washington, DC: Institute of Education Sciences, National Center for Education Statistics. Retrieved from https://nces.ed.gov/statprog/instruct_access_faq.asp#a5

U.S. Department of Education. (2021). NAEP Data Explorer. Washington, DC: Institute of Education Sciences, National Center for Education Statistics. Retrieved February 2021 from https://www.nationsreportcard.gov/ndecore/xplore/nde

United States Census Bureau. (2016). The Hispanic Population in the United States: 2016. Current Population Survey. Retrieved from https://www.census.gov/data/tables/2016/demo/hispanic-origin/2016-cps.html

United States Census Bureau. (2018, October). American Community Survey (ACS): American Community Survey Data. Retrieved from https://www.census.gov/programs-surveys/acs/data.html

United States Department of Treasury. (2019, May). SOI Tax Stats: IRS Data Book. Retrieved from https://www.irs.gov/statistics/soi-tax-stats-irs-data-book

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). Mice: Multivariate Imputation by Chained Equations. *Journal of Statistical Software*, 45(3), 1-67. Retrieved from https://www.jstatstoft.org/v45/i03/.

van Buuren, S. (2018). Flexible Imputation of Missing Data (2nd ed.). Boca Raton, FL: CRC Press. https://stefvanbuuren.name/fimd/sec-algoptions.html

von Davier, M., Gonzalez, E & Mislevy, R.J. (2009). What are plausible values and why are they useful. IERI monograph series, 2, 9-36. Retrieved from http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02_Chapter_01.pdf

Warikoo, N., Sinclair, S., Fei, J., & Jacoby-Senghor, D. (2016). Examining Racial Bias in Education: A New Approach. *Educational Researcher*, 45(9), 508-514.

White, I. R.. Royston, P., & Wood, A. M. (2009). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30, 377–399

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York: NY.

Wickham, H., François, R., Henry, L., & Müller, K. (2019). dplyr: A Grammar of Data Manipulation. R package version 0.8.0.1. https://CRAN.R-project.org/package=dplyr

Willmott, C. & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research, 30(1), 79-82.

Workman, E. (2012). Teacher Expectations of Students. *The Progress of Education Reform*, 13(6). Retrieved from https://www.ecs.org/clearinghouse/01/05/51/10551.pdf

Zuberi, T. (2000). Deracializing Social Statistics: Problems in the Quantification of Race. *The Annals of the American Academy of Political and Social Science*, 568, 172-185.

Appendix A: Implementation of Prediction Techniques in R and Stata

This appendix provides a high-level description of the steps involved in calculating the estimates of mean math achievement for the three techniques: MICE, FH and FLEX CS. The level of detail is intended to provide the reader with a general understanding of the computations involved; however, these descriptions do not provide the information required to reproduce results from this study. To reproduce the results or learn about the statistical computing procedures used for analyses not covered in this appendix, the reader must access the full sets of statistical code provided on author's GitHub page.

Implementation of MICE in R

This section provides a general overview of the programming steps followed to implement the MICE technique in R with the *mice* package (van Buuren & Groothuis-Oudshoorn, 2011). The example state subgroup referenced in the steps, "AL/NHS," represents achievement data for Alabama students whose parents did not finish high school.

1. Import the test sample data and remove the first target value (AL/NHS) with functionality from R's *base* package (R Core Team, 2017).
2. Create an object "vis," a vector of length 10 that defines the visiting sequence (this object is called in later steps during execution of the `mice` function).

```
vis <- c("API", "TP", "BA", "SBA", "HS", "B", "NHS", "H", "AIAN", "EL")
```

   The order is specified in a manner that minimizes the number of initialized values used for predictor variables to begin the iterative process.

3. Using the `mice` function from the *mice* package, create an object "pred_matrix"— a data matrix with 0s along the diagonal and 1s in each off-diagonal cell.

```
for_pred_matrix <- mice([data], maxit = 0, print=F)
```

```
pred_matrix <- for_pred_matrix$pred
```

4. Using data management functionality with R's *base* package, recode off-diagonal values of 1 to 0 for columns (representing subgroup variables) that should not be used to predict rows. The resulting predictor matrix takes the following form—

	E	I	NHS	HS	SBA	BA	W	B	H	API	AIAN	TP	EL	NEL	SWD	NSWD	M	F
E	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NHS	1	0	0	1	1	1	1	1	1	0	1	0	0	1	0	0	0	0
HS	1	1	1	0	1	1	0	0	0	0	0	0	0	1	0	1	1	1
SBA	1	1	1	1	0	1	0	0	0	0	0	0	0	1	0	1	1	1
BA	1	1	0	1	1	0	1	0	0	0	0	0	0	1	0	1	1	1
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B	1	0	1	1	0	0	1	0	1	0	0	0	0	0	0	0	0	0
H	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
API	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
AIAN	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0
TP	0	0	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0
EL	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0
NEL	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SWD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
NSWD	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
M	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

5. Store results from running the `mice` command in an object called "imp1." Set the number of imputations ***m*** to 100, iterations to 15, call visiting sequence order with `visitSequence` option, method to "norm" (Bayesian linear regression), and a seed value to "2019" for reproducibility.

```
imp1 <- mice([data], m = 100, maxit = 15, visitSequence = vis,
        method = "norm", pred = pred_matrix, seed = 2019)
```

6. Store ***m*** imputation values for the first target value of interest (AL/NHS) in an object (vector) called AL_NHS by using the "$imp" command from the *mice* package and subsetting functionality from R's *base* package.

```
AL_NHS <- as.vector(imp1$imp$NHS[1, ])
```

7. Record the mean and standard deviation of ***m*** imputed values, respectively, in objects called "AL_NHS_mean" and "AL_NHS_SE" with the `rowMeans` function from R's *base* package and the `rowSds` function from the *matrixStats* package (Bengtsson, 2018).

```
AL_NHS_mean <- rowMeans(AL_NHS)
AL_NHS_SE <- rowSds(as.matrix(AL_NHS))
```

The value stored in "AL_NHS_mean" represents the predicted value of mean math achievement for this particular subgroup (Alabama students whose parents did not finish high school). It is this value that is compared to the corresponding observed (target) value—a comparison which is employed in the calculations of the weighted Mean Absolute Error (*wMAE*) and coverage statistics for evaluating the performance of MICE technique.

The value stored in "AL_NHS_SE" represents the standard error of the mean math achievement estimate for this particular subgroup (AL/NHS). It is the standard deviation of ***m*** estimates of mean math achievement. These standard errors play an important role in calculating mean math achievement estimates for FLEX CS, the third prediction technique evaluated in this study. The standard errors serve as weights used for calculating precision-weighted estimates of mean math achievement in the FLEX CS approach.

For the remaining target values, repeat steps 1-7, each time withholding a new target value (observed value) from the test sample and returning the previously withheld value. Functionality with the apply family of functions from the *dplyr* package (Wickham et al., 2019) in R is used to automate the repetition of the seven-step cycle.

Implementation of FH in Stata and R

Direct Estimates

The calculation of direct estimates from restricted-use student-level data is implemented through Stata v16.1 and the *svy* package (2019). In the first step, restricted-use data are imported into the software environment and rows are filtered on students in public schools and the subgroup of interest. In the following line of code, `PARED == 1` filters on students whose parents did not finish high school.

```
keep if PUBPRIV == 1 & PARED == 1
```

In the next step, a seed is set to 2019 (the year when dissertation writing began) and a random sample of students equal in size to the median number of students sampled by NAEP in states that are *not reported* by NAEP, for the corresponding subgroup.[50]

```
set seed 2019
sort FIPS15
by FIPS15: sample [censored], count
```

The following line of code sets up analysis of complex survey data. It identifies the location of each student's sampling weight, jackknife replicate weights, and instructs the software to use the jackknife method to calculate standard errors.

```
svyset [pweight = ORIGWT] , jkrweight(SRWT*) vce(jackknife) mse
```

The next lines of code represent a *for loop*, as well as commands that save output generated from executing the *for loop*. The code instructs the software program to compute the mean achievement and standard error for each set of plausible values (20 total) for students from a particular state (Alabama; `FIPS15==1`, in the example code). In addition, the code instructs the program to save the 20 sets of mean and standard error values in a Stata data file (i.e., .dta file).

[50] The sample size is deliberately censored to comply with National Center for Education Statistics reporting policy.

An additional line of code generates mean variance estimates, which are subsequently required to compute pooled variances estimates of the direct estimates of mean achievement.

```
postfile buffer mean_ach stderr using "[pathfile]/results.dta", replace
forvalues i=1(1)20 {
      svy: mean MRPCM`i' if FIPS15 == 1
      mat results = r(table)
      local mean_ach = results[1,1]
      local stderr = results[2,1]
      post buffer (`mean_ach') (`stderr')
}
postclose buffer

clear

use "[pathfile]/results.dta"

gen mean_var = stderr^2

drop stderr

save "[pathfile]/NHS-AL.dta", replace
```

Then, each data file is imported into the R statistical environment and Rubin's rules (1987) are applied to the sets of mean and variance values to calculate pooled estimates for each subgroup. The following sets of code represents user-defined functions written in R that implement Rubin's rules.[51]

```
#write function that pools variance and then takes the sqrt (the se)
pooled_se <- function(x){
      within_var <- mean(x)
      between_var <- var(x)
      sampling_var <- between_var/20
      sqrt(sum(within_var, between_var, sampling_var))
}

#write function that reads .dta file with mean plausible values and
# associated mean variance estimates, and returns a mean and se.

mean_and_se <- function(x) {
       mean_de <- apply(x[,1], 2, mean)
      se_de <- apply(x[,2], 2, pooled_se)
      print(c("x", mean_de, se_de))
}
```

[51] Full documentation of the code used for computing direct estimates is provided on the author's GitHub page.

Calculating the EBLUP

Combining direct and synthetic estimates to calculate EBLUPs is implemented with the *sae* package in R (Molina & Marhuenda, 2015). To calculate the EBLUPs, OLS regression models are successively specified with direct estimates of a subgroup set as values of the response variables. The response variable values all represent direct estimates of some form—for each regression model fit (376 total), one value from the response variable is a direct estimate based on a small sample ($n < 62$) randomly drawn from the restricted-use data and the rest of the response variable values are public-use NAEP-reported direct estimates.

Variables representing factors related to the direct estimates of the subgroup are set as predictor variables. In addition, a vector of values representing the variance estimates associated with the direct estimates are adjoined to the data frame containing the response and predictor variables. The value predicted from the regression fit for the case (state) associated with the replacement estimate is the regression-synthetic estimate for the corresponding state. The `mseFH` function from the *sae* package returns EBLUPs— precision-weighted combination of the direct and synthetic estimates of mean math achievement. The function also returns variance estimates for each EBLUP. The following is a sample of the R code used to compute the EBLUPs of students whose parents did not finish high school.

```
FH_procedure <- function(x){
       FH_df1 <- FH_df
       FH_df1[x, 30] <- FH_df1[x, 2] #30 corresponds with the NR_NHS_Mean column, 2
with the NHS_direct_est column
       FH_df1[x, 31] <- FH_df1[x, 3] #31 corresponds with the NR_NHS_SE column, 3 with
the NHS_se column
       FH_df1 <- filter(FH_df1, NR_NHS_Mean != "NA") #drops non target value rows
       attach(FH_df1)
       mseFH(NR_NHS_Mean ~ B_H_AINA + FER + p_EL + SQI, NR_NHS_SE^2) #line changes per
subgroup
}

NHS_FH_results <- lapply(1:length(FH_df$NR_NHS_Mean), FH_procedure)
```

In the first part of the code, a user-defined function named "FH_procedure" performs the following tasks: takes a row "x", copies the data set with the predictor and response variables required for the analysis, replaces the NAEP-reported estimates of mean achievement and standard errors of row x (representing a state) with the mean and standard error computed from a randomly drawn small sample, removes cases with missing outcome values, and then applies the `mseFH` function from the *sae* package. The code that follows the `mseFH` command represents the regression equation for the NHS subgroup. The last term from the command ("NR_NHS_SE^2") represents the mean variance estimates associated with the values from the response variable. In the last part of the set of code, the `lapply` function is used to iterate the procedure over all rows (i.e., states). The results are saved in an object named "NHS_FH_results" and additional data wrangling tasks (code provided on GitHub) are undertaken to extract the EBLUPs and standard errors associated with the EBLUPs.

Implementation of FLEX CS in R

This section is limited to an explanation of how WPE and NNI subestimates are calculated. The other subestimates that factor into FLEX CS estimates are MICE and FH subestimates. The minor differences in how MICE and FH subestimates are calculated compared to MICE and FH estimates used in the previous two technique are demonstrated on GitHub.

WPE Subestimate Calculation

Calculation of WPE means and variances in R is performed with the assistance of the *dplyr* package (Wickham et al., 2019), in addition to R's *base* package (R Core Team, 2019), following these general steps:[52]

[52] The full set of R code used for implementing WPE is provided on the author's GitHub page.

1. Downloading and importing the "SEDA_geodist_long_NAEP_v21.csv" data file from the SEDA website.
    a. This file contains several variables, including estimates of mean math achievement and standard errors in NAEP-referenced units by grade, year and subject for select subgroups across "geographic districts," which in broad terms represent school districts. The file also includes variables representing counts of students by grade, year and subject for subgroups across districts, which permits the calculation of weights and pooled variance estimates for the WPEs.
2. Sub-setting the file for students in grades 8 in 2015 taking NAEP math, using *dplyr's* piping functionality and `filter` function.
3. Computing sums of students by state across subgroups.
4. Creating variables with dplyr's `mutate` function whose values represent weights by dividing district counts by sums representing total number of students within subgroups by state.
5. By subgroup and state, sum the product of districts' weights and districts' estimates of mean math achievement (as reported in SEDA file) to compute WPE estimates of mean math achievement.
6. By subgroup and state, divide the sum of the products of districts' variances estimates and their counts minus one by the sum of districts' counts minus one to compute WPE variance estimates.

NNI Subestimate Calculation

To begin implementation of the NNI subestimate approach in R, variables are standardized using the `scale` function from R's *base* package (R Core Team, 2019).

Euclidean distances are then computed using the `dist` function from R's *stats* package (R Core Team, 2019), which generates an *n*-by-*n* distance matrix, where *n* corresponds to the number of cases (observations) from the data frame used for analysis. Each element (cell) of the matrix is a Euclidean distance indicating the degree of dissimilarity between corresponding cases, where larger distance values indicate less similarity.

In this dissertation, cases correspond to states and thus implementation of the `dist` function results in a 50-by-50 distance matrix, with cell values that indicate the degree to which corresponding states are similar (dissimilar) based on their values on the four separate data variables. A state has a sibling, and thus a donor, if its nearest neighbor (i.e., most similar state) has a Euclidean distance of less than 0.4 standard deviations (< 0.4 SD).

The standard deviation is computed by taking the square root of the variance of all dissimilarity values between states—that is, the standard deviation of 1225 values representing the Euclidean distance between pairs of states. Hence, the distance criterion used for establishing whether states are siblings represents a relative distance as opposed to an absolute distance.

Appendix B: Subgroup Tables of Technique-produced Estimates of Mean Math Achievement

Table B.1. Estimates for students whose parents did not finish high school by state and technique, including NAEP-reported estimates

Did not finish high school (NHS)	NAEP Reported	MICE	FH	FLEX CS
AL	254 (2.5)	252 (4.2)	259 (3.3)	261 (4.0)
AK	--	--	--	--
AZ	269 (2.3)	270 (4.3)	267 (2.0)	267 (2.0)
AR	266 (2.4)	263 (3.5)	265 (2.1)	265 (2.1)
CA	261 (2.0)	259 (3.7)	263 (1.9)	263 (1.9)
CO	266 (2.6)	265 (3.4)	267 (2.2)	267 (2.2)
CT	254 (4.2)	260 (4.3)	261 (3.0)	263 (5.2)
DE	264 (2.8)	264 (3.9)	265 (2.3)	265 (2.2)
FL	264 (2.5)	264 (3.6)	264 (2.2)	264 (2.2)
GA	269 (2.4)	266 (3.4)	267 (2.1)	267 (2.1)
HI	270 (4.4)	265 (3.7)	270 (3.2)	269 (3.0)
ID	262 (2.6)	264 (3.7)	262 (2.3)	262 (2.2)
IL	270 (3.1)	266 (3.3)	268 (2.5)	268 (2.5)
IN	268 (3.1)	268 (3.7)	266 (2.5)	266 (2.4)
IA	261 (3.5)	264 (3.7)	263 (2.7)	264 (4.0)
KS	268 (4.1)	268 (3.7)	267 (3.0)	264 (3.9)
KY	260 (2.6)	263 (3.6)	261 (2.2)	261 (2.2)
LA	258 (2.5)	262 (4.0)	259 (2.2)	260 (2.1)
ME	267 (4.4)	264 (4.1)	266 (3.0)	266 (2.9)
MD	265 (3.4)	265 (3.7)	266 (2.8)	267 (2.7)
MA	267 (4.6)	273 (3.5)	268 (3.2)	268 (3.1)
MI	261 (3.7)	260 (4.0)	263 (2.7)	260 (4.8)
MN	275 (3.3)	268 (3.7)	272 (2.6)	272 (2.6)
MS	258 (3.0)	267 (4.0)	259 (2.5)	259 (2.4)
MO	257 (2.9)	262 (3.8)	259 (2.4)	260 (3.4)
MT	272 (3.7)	271 (3.6)	268 (2.8)	268 (2.7)
NE	262 (2.7)	265 (3.9)	263 (2.3)	264 (4.7)
NV	263 (2.0)	261 (3.5)	264 (1.9)	264 (1.9)
NH	269 (4.4)	266 (3.8)	268 (3.2)	267 (3.1)
NJ	267 (4.5)	267 (4.3)	267 (3.2)	262 (10.1)
NM	261 (2.2)	263 (3.5)	262 (2.0)	262 (2.0)
NY	267 (3.0)	266 (3.7)	267 (2.5)	267 (2.4)
NC	264 (2.5)	266 (3.5)	264 (2.1)	264 (2.1)
ND	266 (3.4)	268 (3.5)	265 (2.7)	264 (4.5)
OH	259 (4.6)	263 (3.9)	262 (3.1)	262 (3.0)
OK	263 (3.0)	259 (3.8)	263 (2.4)	259 (6.6)
OR	268 (2.4)	263 (4.0)	267 (2.1)	267 (2.1)
PA	261 (3.4)	259 (3.9)	262 (2.6)	263 (3.2)
RI	268 (2.5)	260 (3.7)	267 (2.2)	267 (2.1)
SC	271 (3.6)	264 (3.3)	267 (2.7)	267 (2.7)
SD	265 (4.1)	267 (3.5)	264 (2.9)	264 (2.8)
TN	265 (3.1)	260 (4.0)	264 (2.5)	264 (2.5)
TX	272 (1.9)	272 (4.3)	271 (1.7)	271 (1.7)
UT	--	--	--	--
VT	266 (3.6)	269 (4.9)	266 (2.7)	266 (2.6)
VA	268 (3.2)	269 (4.1)	267 (2.6)	267 (2.5)
WA	266 (3.0)	266 (3.5)	266 (2.4)	266 (2.4)
WV	255 (2.9)	260 (3.8)	258 (2.5)	258 (2.4)
WI	263 (3.8)	266 (4.4)	264 (2.8)	263 (3.7)
WY	272 (2.8)	267 (3.4)	270 (2.4)	269 (2.3)

Table B.2. Estimates for students whose parents graduated from high school by state and technique, including NAEP-reported estimates

Graduated high school (HS)	NAEP Reported	MICE	FH	FLEX CS
AL	252 (2.3)	254 (2.5)	259 (3.3)	261 (6.2)
AK	--	--	--	--
AZ	271 (2.7)	271 (2.3)	268 (2.3)	270 (2.6)
AR	262 (2.4)	264 (2.2)	263 (2.1)	264 (2.4)
CA	263 (2.3)	264 (2.5)	265 (2.1)	263 (2.3)
CO	270 (2.3)	271 (2.5)	270 (2.0)	270 (2.3)
CT	265 (2.1)	266 (2.8)	266 (1.9)	268 (4.7)
DE	267 (2.0)	266 (2.3)	267 (1.8)	267 (2.3)
FL	266 (2.1)	264 (2.1)	265 (1.9)	265 (2.3)
GA	264 (1.9)	266 (2.5)	263 (1.7)	265 (2.5)
HI	264 (1.7)	269 (2.3)	266 (1.6)	266 (2.9)
ID	267 (2.5)	272 (2.2)	267 (2.2)	269 (3.2)
IL	268 (2.0)	269 (2.2)	268 (1.8)	268 (2.0)
IN	274 (1.7)	273 (2.2)	273 (1.6)	273 (1.9)
IA	272 (2.4)	272 (2.3)	272 (2.1)	272 (2.6)
KS	272 (2.4)	271 (2.3)	272 (2.1)	270 (3.2)
KY	268 (1.6)	265 (2.2)	268 (1.5)	267 (2.8)
LA	259 (2.2)	258 (2.3)	260 (2.0)	259 (2.8)
ME	272 (1.9)	272 (2.5)	272 (1.8)	272 (2.0)
MD	263 (1.9)	267 (2.2)	264 (1.7)	265 (2.2)
MA	277 (2.7)	281 (2.8)	276 (2.3)	278 (4.1)
MI	263 (2.1)	265 (2.3)	264 (1.9)	266 (3.1)
MN	278 (2.9)	277 (2.4)	276 (2.4)	276 (2.3)
MS	259 (2.0)	259 (2.7)	259 (1.9)	260 (2.1)
MO	268 (2.0)	266 (2.4)	268 (1.8)	266 (3.3)
MT	270 (2.3)	276 (2.2)	270 (2.1)	273 (4.6)
NE	267 (2.1)	271 (2.2)	268 (1.9)	270 (3.2)
NV	266 (1.5)	263 (2.5)	266 (1.4)	265 (2.4)
NH	278 (1.9)	277 (2.4)	277 (1.8)	277 (2.1)
NJ	273 (2.2)	273 (2.5)	272 (2.0)	270 (5.1)
NM	260 (1.6)	263 (2.4)	260 (1.5)	261 (3.2)
NY	269 (2.5)	269 (2.2)	269 (2.2)	269 (2.2)
NC	267 (2.1)	267 (2.3)	267 (1.9)	267 (2.1)
ND	274 (2.3)	272 (2.3)	273 (2.0)	272 (2.3)
OH	272 (2.1)	270 (2.2)	271 (1.9)	270 (2.2)
OK	265 (1.8)	264 (2.3)	265 (1.7)	261 (7.2)
OR	271 (1.9)	270 (2.2)	271 (1.7)	270 (2.2)
PA	265 (3.3)	268 (2.2)	266 (2.6)	268 (3.5)
RI	269 (1.8)	266 (2.1)	269 (1.7)	268 (2.9)
SC	263 (2.3)	263 (2.2)	263 (2.1)	263 (2.1)
SD	269 (1.9)	271 (2.3)	269 (1.7)	270 (2.1)
TN	266 (2.0)	265 (2.2)	266 (1.8)	265 (2.2)
TX	275 (1.8)	272 (2.2)	273 (1.7)	273 (2.3)
UT	--	--	--	--
VT	277 (2.0)	277 (2.6)	276 (1.8)	276 (2.2)
VA	272 (2.0)	271 (2.4)	271 (1.8)	271 (1.9)
WA	272 (2.6)	272 (2.3)	272 (2.2)	272 (2.2)
WV	261 (1.6)	262 (2.3)	263 (1.5)	262 (1.9)
WI	271 (2.4)	270 (2.2)	271 (2.1)	270 (4.3)
WY	272 (1.7)	273 (2.4)	272 (1.6)	272 (1.9)

Table B.3. Estimates for students whose parents have some education after high school by state and technique, including NAEP-reported estimates

Some education after high school (SBA)	NAEP Reported	MICE	FH	FLEX CS
AL	271 (1.9)	276 (3.1)	274 (3.0)	275 (4.6)
AK	--	--	--	--
AZ	285 (1.9)	283 (1.4)	284 (1.7)	284 (2.3)
AR	280 (1.9)	280 (2.5)	280 (1.7)	280 (2.2)
CA	278 (2.2)	281 (2.8)	279 (2.0)	280 (2.5)
CO	284 (2.3)	284 (2.0)	284 (2.0)	284 (2.3)
CT	275 (2.4)	283 (1.9)	278 (2.1)	284 (7.0)
DE	280 (1.8)	280 (2.9)	280 (1.6)	280 (2.1)
FL	280 (2.0)	280 (2.3)	280 (1.8)	280 (2.3)
GA	281 (2.0)	282 (3.0)	280 (1.8)	280 (2.2)
HI	284 (1.6)	284 (2.8)	285 (1.5)	285 (2.3)
ID	286 (1.9)	285 (2.6)	285 (1.7)	285 (2.2)
IL	282 (1.7)	283 (1.7)	282 (1.6)	282 (2.2)
IN	288 (2.0)	287 (2.0)	287 (1.8)	287 (2.2)
IA	283 (2.0)	285 (2.2)	283 (1.8)	285 (2.4)
KS	282 (1.8)	285 (2.4)	283 (1.7)	285 (3.5)
KY	281 (1.6)	281 (2.5)	281 (1.5)	281 (2.2)
LA	270 (1.9)	276 (1.9)	271 (1.7)	273 (3.0)
ME	283 (1.8)	285 (2.3)	284 (1.7)	284 (2.6)
MD	282 (2.0)	282 (2.4)	282 (1.8)	282 (2.2)
MA	293 (2.2)	289 (1.9)	291 (2.0)	293 (4.0)
MI	275 (2.2)	279 (2.5)	277 (1.9)	280 (4.6)
MN	290 (1.8)	289 (2.0)	289 (1.6)	290 (2.7)
MS	279 (2.5)	277 (2.2)	278 (2.2)	277 (2.8)
MO	285 (2.0)	279 (2.2)	284 (1.8)	280 (5.1)
MT	286 (1.8)	288 (2.0)	286 (1.7)	286 (3.0)
NE	288 (2.3)	285 (2.3)	287 (2.0)	284 (3.1)
NV	283 (1.7)	278 (2.0)	282 (1.6)	281 (3.4)
NH	290 (1.7)	288 (1.9)	290 (1.6)	290 (2.2)
NJ	291 (2.1)	287 (1.8)	289 (1.9)	284 (7.8)
NM	277 (1.9)	278 (2.8)	277 (1.8)	277 (2.3)
NY	283 (1.9)	283 (1.6)	283 (1.8)	283 (2.2)
NC	281 (1.8)	283 (1.4)	281 (1.6)	282 (2.3)
ND	286 (1.6)	286 (2.6)	286 (1.5)	285 (2.8)
OH	284 (2.0)	285 (2.4)	284 (1.8)	284 (2.2)
OK	277 (2.1)	277 (2.7)	278 (1.8)	275 (4.6)
OR	283 (2.1)	284 (2.0)	283 (1.9)	284 (2.5)
PA	284 (2.3)	283 (1.0)	284 (2.0)	285 (4.0)
RI	285 (2.0)	282 (2.0)	285 (1.8)	284 (3.0)
SC	278 (2.2)	279 (2.7)	278 (1.9)	279 (2.2)
SD	287 (1.9)	285 (2.5)	286 (1.7)	286 (2.1)
TN	281 (2.0)	279 (2.8)	281 (1.8)	280 (2.5)
TX	284 (2.0)	286 (2.6)	283 (1.8)	284 (3.3)
UT	--	--	--	--
VT	288 (2.5)	288 (2.4)	288 (2.1)	288 (2.5)
VA	281 (1.9)	285 (2.2)	282 (1.7)	283 (2.9)
WA	288 (1.8)	286 (2.7)	288 (1.6)	287 (2.6)
WV	275 (1.8)	277 (2.5)	276 (1.6)	276 (2.1)
WI	289 (2.1)	286 (2.4)	288 (1.9)	286 (2.7)
WY	285 (1.8)	287 (2.4)	285 (1.7)	286 (2.6)

Table B.4. Estimates for students whose parents graduated from college by state and technique, including NAEP-reported estimates

Graduated from college (BA)	NAEP Reported	MICE	FH	FLEX CS
AL	276 (1.4)	280 (2.9)	276 (3.7)	281 (5.6)
AK	--	--	--	--
AZ	297 (2.0)	294 (1.5)	295 (1.9)	294 (2.2)
AR	283 (1.6)	285 (1.6)	284 (1.5)	285 (1.9)
CA	293 (1.7)	289 (1.6)	293 (1.6)	292 (3.0)
CO	299 (1.6)	299 (1.9)	299 (1.6)	299 (1.9)
CT	297 (1.4)	296 (1.4)	297 (1.3)	299 (4.4)
DE	290 (1.1)	290 (1.7)	291 (1.1)	290 (1.8)
FL	285 (1.6)	285 (2.3)	286 (1.5)	286 (1.7)
GA	289 (1.5)	291 (1.8)	289 (1.5)	289 (2.2)
HI	289 (1.2)	290 (2.0)	290 (1.2)	290 (1.5)
ID	294 (1.2)	293 (1.7)	294 (1.1)	294 (1.6)
IL	293 (1.8)	294 (1.4)	293 (1.7)	293 (1.7)
IN	297 (1.4)	297 (1.0)	297 (1.4)	297 (1.7)
IA	296 (1.3)	297 (0.8)	296 (1.3)	296 (1.5)
KS	293 (1.3)	296 (1.4)	293 (1.2)	296 (2.7)
KY	288 (1.1)	285 (1.5)	288 (1.1)	287 (1.8)
LA	277 (1.6)	280 (3.5)	278 (1.5)	278 (1.7)
ME	295 (1.0)	296 (1.7)	295 (1.0)	295 (1.4)
MD	295 (1.6)	294 (1.8)	295 (1.6)	295 (2.2)
MA	308 (1.4)	302 (1.9)	308 (1.3)	308 (2.1)
MI	288 (1.5)	289 (1.7)	288 (1.5)	289 (2.1)
MN	304 (1.2)	303 (3.2)	303 (1.2)	304 (2.8)
MS	277 (1.5)	282 (2.9)	277 (1.5)	279 (3.4)
MO	291 (1.4)	290 (1.6)	291 (1.4)	290 (2.2)
MT	296 (1.0)	297 (1.2)	296 (1.0)	296 (1.8)
NE	298 (1.0)	296 (0.9)	298 (1.0)	296 (2.8)
NV	288 (1.3)	288 (2.4)	288 (1.3)	288 (1.6)
NH	304 (1.0)	302 (2.8)	304 (1.0)	304 (1.4)
NJ	303 (1.6)	303 (2.9)	303 (1.5)	301 (4.2)
NM	283 (1.4)	285 (1.8)	284 (1.4)	284 (1.8)
NY	290 (1.6)	292 (2.0)	290 (1.6)	291 (1.7)
NC	294 (2.0)	294 (1.8)	294 (1.9)	293 (1.8)
ND	296 (0.9)	297 (1.0)	296 (0.8)	296 (1.8)
OH	296 (1.4)	296 (1.4)	295 (1.4)	295 (1.6)
OK	284 (1.7)	285 (1.7)	285 (1.6)	281 (5.3)
OR	295 (1.6)	293 (1.4)	295 (1.5)	294 (2.1)
PA	297 (1.6)	294 (1.7)	297 (1.5)	297 (2.8)
RI	293 (1.0)	293 (1.7)	293 (1.0)	293 (1.4)
SC	284 (1.3)	286 (2.0)	284 (1.2)	285 (1.8)
SD	292 (1.1)	295 (1.4)	292 (1.1)	293 (1.9)
TN	291 (2.2)	288 (1.6)	290 (2.0)	289 (2.6)
TX	296 (1.6)	295 (1.6)	296 (1.5)	296 (1.8)
UT	--	--	--	--
VT	301 (1.1)	299 (1.5)	301 (1.1)	301 (1.9)
VA	298 (1.6)	298 (1.5)	298 (1.5)	299 (1.7)
WA	300 (1.4)	298 (1.6)	299 (1.4)	299 (1.5)
WV	280 (1.4)	279 (3.6)	281 (1.4)	280 (2.0)
WI	299 (1.2)	299 (1.5)	298 (1.2)	298 (1.9)
WY	297 (1.0)	296 (1.4)	297 (1.0)	297 (1.8)

Table B.5. Estimates for Black students by state and technique, including NAEP-reported estimates

Black (B)	NAEP Reported	MICE	FH	FLEX CS
AL	248 (1.8)	249 (4.8)	254 (2.7)	256 (6.4)
AK	269 (4.5)	264 (4.5)	263 (3.7)	261 (3.8)
AZ	269 (4.1)	263 (4.7)	263 (3.5)	264 (5.5)
AR	255 (2.2)	258 (4.7)	254 (3.2)	254 (3.2)
CA	260 (3.3)	255 (4.5)	263 (4.2)	259 (7.5)
CO	260 (4.9)	261 (4.4)	265 (3.8)	265 (3.8)
CT	256 (2.9)	252 (5.4)	261 (3.5)	263 (7.3)
DE	263 (1.3)	259 (4.2)	260 (3.4)	259 (3.2)
FL	258 (2.2)	259 (2.3)	260 (4.0)	260 (4.0)
GA	264 (1.5)	260 (4.6)	258 (3.5)	262 (3.6)
HI	--	--	--	--
ID	--	--	--	--
IL	261 (2.4)	262 (4.9)	260 (2.1)	260 (2.1)
IN	257 (3.2)	263 (4.9)	257 (2.5)	258 (6.1)
IA	254 (3.0)	259 (4.8)	255 (2.5)	256 (8.5)
KS	263 (3.8)	262 (5.0)	261 (2.9)	260 (6.4)
KY	257 (2.3)	258 (5.4)	257 (2.0)	256 (2.5)
LA	255 (1.4)	256 (4.9)	255 (1.3)	255 (2.3)
ME	--	--	--	--
MD	263 (1.3)	260 (5.7)	263 (1.3)	264 (3.3)
MA	268 (3.6)	264 (5.7)	267 (2.9)	269 (6.5)
MI	251 (2.2)	257 (5.0)	252 (1.9)	254 (5.6)
MN	262 (2.7)	265 (5.9)	262 (2.4)	263 (3.9)
MS	257 (1.7)	257 (5.2)	257 (1.6)	257 (2.7)
MO	258 (2.7)	256 (5.5)	258 (2.2)	254 (5.7)
MT	--	--	--	--
NE	254 (3.6)	262 (4.9)	256 (2.7)	257 (5.8)
NV	256 (2.5)	257 (4.6)	257 (2.2)	257 (2.2)
NH	--	--	--	--
NJ	269 (3.0)	261 (4.6)	266 (2.5)	262 (7.3)
NM	--	--	--	--
NY	264 (2.9)	260 (4.5)	263 (2.5)	263 (2.4)
NC	263 (2.0)	262 (4.3)	262 (1.8)	262 (2.4)
ND	263 (4.7)	263 (4.6)	261 (3.5)	257 (6.1)
OH	259 (3.1)	257 (4.6)	258 (2.5)	258 (2.5)
OK	260 (2.7)	254 (4.5)	259 (2.3)	253 (6.7)
OR	--	--	--	--
PA	253 (2.3)	256 (5.2)	254 (2.0)	254 (5.2)
RI	258 (3.1)	255 (5.2)	259 (2.6)	259 (2.6)
SC	256 (1.9)	261 (5.0)	257 (1.7)	257 (2.7)
SD	--	--	--	--
TN	253 (3.0)	260 (4.0)	255 (2.4)	258 (4.9)
TX	267 (2.9)	266 (4.3)	265 (2.4)	264 (2.4)
UT	--	--	--	--
VT	--	--	--	--
VA	265 (1.8)	264 (5.2)	265 (1.7)	265 (1.7)
WA	257 (3.4)	260 (4.8)	260 (2.6)	260 (2.6)
WV	256 (3.1)	248 (5.6)	254 (2.7)	254 (5.2)
WI	249 (4.2)	261 (4.0)	253 (3.0)	253 (2.6)
WY	--	--	--	--

Table B.6. Estimates for Hispanic students by state and technique, including NAEP-reported estimates

Hispanic (H)	NAEP Reported	MICE	FH	FLEX CS
AL	260 (3.2)	263 (4.4)	265 (3.4)	264 (6.1)
AK	279 (2.9)	270 (3.6)	272 (4.0)	273 (4.2)
AZ	273 (1.2)	272 (3.7)	268 (3.6)	271 (4.2)
AR	269 (2.8)	270 (3.8)	270 (3.7)	269 (3.4)
CA	263 (1.5)	268 (3.7)	271 (4.5)	265 (5.5)
CO	269 (1.6)	271 (3.7)	267 (3.9)	268 (3.5)
CT	261 (2.3)	265 (4.3)	270 (3.9)	270 (7.4)
DE	270 (2.2)	271 (4.3)	271 (3.5)	273 (6.8)
FL	272 (1.4)	268 (3.6)	267 (4.4)	271 (4.9)
GA	270 (2.1)	272 (3.6)	271 (3.7)	272 (3.4)
HI	271 (2.9)	270 (4.1)	271 (4.6)	273 (3.7)
ID	264 (2.2)	269 (4.2)	268 (3.8)	267 (5.0)
IL	273 (1.4)	271 (3.6)	271 (4.2)	271 (3.9)
IN	271 (3.1)	272 (3.8)	271 (3.9)	272 (5.5)
IA	269 (2.1)	267 (3.8)	272 (3.8)	272 (6.4)
KS	274 (2.8)	275 (4.1)	269 (3.9)	267 (3.6)
KY	274 (2.9)	267 (4.1)	270 (4.1)	268 (4.7)
LA	271 (3.7)	266 (4.6)	269 (4.0)	266 (5.4)
ME	--	--	--	--
MD	273 (2.4)	271 (3.2)	272 (2.2)	271 (3.4)
MA	271 (3.1)	274 (3.3)	270 (2.6)	271 (2.5)
MI	269 (4.0)	268 (3.5)	270 (2.9)	269 (4.8)
MN	272 (3.2)	274 (3.4)	271 (2.6)	271 (3.8)
MS	269 (4.9)	267 (4.1)	269 (3.3)	271 (7.7)
MO	270 (3.5)	266 (4.1)	270 (2.7)	270 (3.4)
MT	275 (4.9)	272 (4.6)	272 (3.3)	272 (3.1)
NE	266 (2.1)	266 (4.2)	267 (1.9)	269 (4.4)
NV	266 (1.1)	268 (3.1)	266 (1.1)	266 (1.1)
NH	270 (4.3)	272 (4.9)	270 (3.2)	270 (3.0)
NJ	272 (2.0)	274 (4.4)	272 (1.8)	268 (7.6)
NM	266 (1.1)	267 (3.5)	267 (1.0)	266 (4.3)
NY	268 (1.7)	271 (3.2)	269 (1.6)	269 (1.6)
NC	273 (2.5)	271 (3.2)	272 (2.2)	272 (2.7)
ND	276 (3.3)	271 (3.6)	274 (2.7)	272 (4.5)
OH	266 (8.8)	267 (3.2)	269 (3.8)	269 (3.6)
OK	266 (2.9)	269 (3.1)	267 (2.4)	265 (4.6)
OR	266 (1.7)	271 (4.4)	267 (1.6)	267 (3.0)
PA	261 (3.7)	265 (3.2)	265 (2.9)	266 (8.1)
RI	265 (1.1)	268 (3.2)	265 (1.1)	265 (1.1)
SC	272 (4.3)	274 (3.6)	270 (3.0)	270 (3.6)
SD	272 (4.5)	270 (3.9)	270 (3.1)	270 (4.2)
TN	273 (4.0)	267 (3.9)	271 (2.9)	271 (3.3)
TX	277 (1.4)	276 (3.4)	276 (1.3)	276 (1.3)
UT	262 (2.6)	267 (3.9)	264 (2.2)	263 (3.2)
VT	--	--	--	--
VA	279 (2.4)	274 (3.6)	277 (2.1)	276 (2.0)
WA	269 (2.5)	269 (3.8)	269 (2.2)	270 (2.2)
WV	--	--	--	--
WI	271 (2.6)	267 (4.0)	271 (2.3)	269 (6.5)
WY	273 (1.9)	271 (4.5)	273 (1.8)	273 (3.2)

Table B.7. Estimates for Asian Pacific Islander students by state and technique, including NAEP-reported estimates

Asian / Pacific Islander (API)	NAEP Reported	MICE	FH	FLEX CS
AL	--	--	--	--
AK	275 (1.9)	303 (9.9)	275 (1.9)	276 (4.6)
AZ	305 (5.1)	304 (11.2)	307 (4.7)	311 (8.3)
AR	--	--	--	--
CA	303 (3.6)	293 (11.8)	303 (3.5)	305 (5.3)
CO	303 (5.5)	306 (10.6)	307 (5.0)	307 (5.0)
CT	310 (5.5)	305 (10.5)	311 (4.9)	323 (12.7)
DE	317 (4.0)	296 (11.0)	315 (3.7)	314 (6.0)
FL	297 (3.4)	302 (11.0)	298 (3.3)	301 (6.1)
GA	317 (6.2)	303 (10.6)	314 (5.3)	314 (4.4)
HI	279 (1.0)	295 (11.5)	278 (1.0)	278 (1.1)
ID	--	--	--	--
IL	309 (4.9)	298 (10.6)	308 (4.4)	313 (9.2)
IN	--	--	--	--
IA	291 (5.0)	306 (12.7)	293 (4.5)	294 (6.7)
KS	301 (5.4)	305 (13.0)	302 (4.8)	305 (6.2)
KY	304 (4.7)	303 (13.3)	303 (4.3)	303 (5.2)
LA	--	--	--	--
ME	--	--	--	--
MD	314 (3.6)	306 (11.5)	314 (3.4)	316 (4.0)
MA	324 (4.2)	323 (13.5)	322 (3.9)	323 (14.3)
MI	313 (4.6)	298 (11.6)	311 (4.2)	314 (10.4)
MN	293 (4.0)	317 (11.3)	296 (3.8)	297 (4.4)
MS	--	--	--	--
MO	--	--	--	--
MT	--	--	--	--
NE	--	--	--	--
NV	294 (3.4)	292 (12.1)	294 (3.3)	294 (3.3)
NH	312 (6.9)	317 (11.5)	314 (5.9)	314 (5.7)
NJ	331 (3.6)	314 (10.9)	329 (3.4)	323 (14.1)
NM	--	--	--	--
NY	298 (3.6)	307 (12.5)	299 (3.5)	299 (3.4)
NC	309 (6.2)	306 (13.0)	310 (5.3)	314 (7.5)
ND	--	--	--	--
OH	305 (24.9)	310 (12.0)	306 (8.2)	306 (8.1)
OK	--	--	--	--
OR	304 (5.5)	305 (11.8)	305 (4.9)	307 (5.4)
PA	316 (6.0)	306 (11.7)	313 (5.1)	306 (11.8)
RI	299 (3.9)	303 (11.8)	300 (3.6)	300 (3.6)
SC	--	--	--	--
SD	--	--	--	--
TN	--	--	--	--
TX	312 (3.2)	307 (11.0)	310 (3.1)	310 (3.1)
UT	--	--	--	--
VT	--	--	--	--
VA	317 (4.1)	312 (10.1)	316 (3.8)	316 (3.8)
WA	309 (3.4)	302 (13.5)	308 (3.2)	308 (3.2)
WV	--	--	--	--
WI	295 (4.7)	311 (10.2)	297 (4.3)	302 (12.6)
WY	--	--	--	--

Table B.8. Estimates for American Indian/Alaskan Native students by state and technique, including NAEP-reported estimates

American Indian / Alaskan Native (AIAN)	NAEP Reported	MICE	FH	FLEX CS
AL	--	--	--	--
AK	257 (2.6)	260 (8.4)	257 (2.6)	257 (2.6)
AZ	260 (4.4)	256 (7.3)	262 (3.8)	262 (3.6)
AR	--	--	--	--
CA	--	--	--	--
CO	--	--	--	--
CT	--	--	--	--
DE	--	--	--	--
FL	--	--	--	--
GA	--	--	--	--
HI	--	--	--	--
ID	--	--	--	--
IL	--	--	--	--
IN	--	--	--	--
IA	--	--	--	--
KS	--	--	--	--
KY	--	--	--	--
LA	--	--	--	--
ME	--	--	--	--
MD	--	--	--	--
MA	--	--	--	--
MI	--	--	--	--
MN	261 (5.8)	263 (7.1)	257 (4.8)	257 (4.2)
MS	--	--	--	--
MO	--	--	--	--
MT	256 (2.9)	262 (6.3)	257 (2.7)	258 (2.6)
NE	--	--	--	--
NV	--	--	--	--
NH	--	--	--	--
NJ	--	--	--	--
NM	259 (2.8)	260 (5.8)	260 (2.7)	260 (2.7)
NY	--	--	--	--
NC	261 (4.7)	261 (5.3)	261 (3.6)	261 (3.4)
ND	260 (3.1)	262 (6.2)	260 (2.8)	260 (2.7)
OH	--	--	--	--
OK	269 (1.8)	259 (5.0)	268 (1.8)	268 (1.8)
OR	--	--	--	--
PA	--	--	--	--
RI	--	--	--	--
SC	--	--	--	--
SD	260 (2.9)	260 (7.5)	260 (2.6)	260 (2.6)
TN	--	--	--	--
TX	--	--	--	--
UT	240 (9.0)	260 (5.5)	256 (4.4)	256 (3.8)
VT	--	--	--	--
VA	--	--	--	--
WA	264 (7.0)	259 (5.0)	259 (4.7)	260 (3.7)
WV	--	--	--	--
WI	274 (7.0)	261 (4.8)	262 (3.9)	262 (3.7)
WY	251 (4.0)	261 (6.1)	255 (3.6)	255 (3.2)

Table B.9. Estimates for students who identify as more than one race by state and technique, including NAEP-reported estimates

More than once race (TP)	NAEP Reported	MICE	FH	FLEX CS
AL	--	--	--	--
AK	285 (2.6)	285 (6.0)	284 (2.5)	284 (2.5)
AZ	--	--	--	--
AR	--	--	--	--
CA	289 (9.7)	282 (5.5)	288 (3.5)	288 (3.5)
CO	290 (4.8)	287 (6.1)	291 (3.3)	291 (3.3)
CT	--	--	--	--
DE	--	--	--	--
FL	282 (3.5)	277 (6.0)	282 (2.7)	282 (2.7)
GA	277 (5.3)	285 (6.3)	281 (2.9)	281 (2.9)
HI	285 (2.6)	277 (5.4)	286 (2.7)	286 (2.7)
ID	--	--	--	--
IL	--	--	--	--
IN	281 (4.6)	284 (5.5)	277 (2.8)	277 (2.8)
IA	283 (5.6)	280 (5.7)	278 (2.8)	278 (2.8)
KS	278 (3.1)	280 (5.6)	279 (2.7)	280 (5.7)
KY	266 (5.9)	275 (6.3)	272 (3.2)	272 (3.2)
LA	--	--	--	--
ME	--	--	--	--
MD	290 (3.9)	286 (5.9)	290 (3.3)	290 (3.3)
MA	--	--	--	--
MI	--	--	--	--
MN	284 (5.0)	293 (6.2)	285 (2.7)	285 (2.7)
MS	--	--	--	--
MO	--	--	--	--
MT	287 (4.1)	281 (5.8)	283 (2.8)	283 (2.8)
NE	285 (5.3)	284 (5.9)	279 (2.7)	279 (3.0)
NV	281 (3.3)	279 (5.9)	280 (2.9)	280 (2.9)
NH	--	--	--	--
NJ	--	--	--	--
NM	--	--	--	--
NY	--	--	--	--
NC	274 (4.4)	285 (5.8)	281 (2.7)	281 (2.7)
ND	--	--	--	--
OH	280 (3.6)	281 (5.3)	278 (2.8)	278 (2.8)
OK	273 (3.8)	273 (5.9)	277 (2.7)	277 (2.7)
OR	281 (4.2)	279 (5.3)	283 (2.9)	283 (2.7)
PA	274 (4.9)	283 (5.2)	278 (3.0)	278 (3.0)
RI	274 (3.3)	281 (5.2)	280 (2.9)	280 (2.9)
SC	--	--	--	--
SD	--	--	--	--
TN	--	--	--	--
TX	293 (6.1)	286 (7.4)	284 (3.2)	284 (3.2)
UT	--	--	--	--
VT	--	--	--	--
VA	293 (3.3)	284 (5.5)	290 (2.7)	290 (2.7)
WA	285 (3.7)	283 (5.9)	285 (2.7)	285 (2.7)
WV	--	--	--	--
WI	--	--	--	--
WY	--	--	--	--

Table B.10. Estimates for English learners by state and technique, including NAEP-reported estimates

English learners (EL)	NAEP Reported	MICE	FH	FLEX CS
AL	--	--	--	--
AK	236 (2.8)	256 (9.6)	237 (2.7)	237 (2.7)
AZ	234 (4.8)	248 (7.7)	237 (4.3)	237 (4.3)
AR	255 (3.2)	243 (10.3)	254 (3.0)	254 (3.0)
CA	238 (1.9)	236 (10.9)	238 (1.9)	239 (1.9)
CO	250 (3.1)	243 (9.6)	250 (2.9)	250 (2.9)
CT	233 (4.3)	236 (11.1)	235 (4.0)	235 (4.0)
DE	--	--	--	--
FL	240 (3.2)	249 (9.2)	241 (3.1)	241 (3.1)
GA	242 (5.2)	245 (9.8)	244 (4.6)	244 (4.6)
HI	239 (2.7)	251 (10.8)	239 (2.7)	239 (2.7)
ID	--	--	--	--
IL	247 (3.4)	250 (10.0)	247 (3.2)	247 (3.2)
IN	260 (5.3)	246 (8.9)	258 (4.7)	257 (4.7)
IA	246 (4.8)	246 (10.6)	247 (4.3)	247 (4.3)
KS	266 (4.0)	249 (8.6)	263 (3.7)	263 (3.7)
KY	--	--	--	--
LA	--	--	--	--
ME	--	--	--	--
MD	247 (4.2)	249 (9.7)	246 (3.9)	247 (3.9)
MA	251 (3.8)	245 (9.7)	250 (3.6)	250 (3.6)
MI	258 (4.8)	244 (9.9)	256 (4.3)	256 (4.3)
MN	252 (3.8)	247 (6.8)	251 (3.5)	251 (3.5)
MS	--	--	--	--
MO	--	--	--	--
MT	--	--	--	--
NE	--	--	--	--
NV	246 (1.7)	242 (9.9)	246 (1.7)	246 (1.7)
NH	--	--	--	--
NJ	--	--	--	--
NM	240 (1.8)	243 (7.8)	240 (1.8)	240 (1.8)
NY	242 (3.8)	245 (9.9)	243 (3.6)	244 (3.6)
NC	247 (4.6)	249 (8.1)	247 (4.1)	247 (4.1)
ND	--	--	--	--
OH	235 (16.8)	242 (10.0)	248 (8.5)	247 (8.2)
OK	245 (4.2)	249 (8.5)	246 (3.9)	245 (3.8)
OR	--	--	--	--
PA	234 (5.7)	237 (10.2)	239 (5.0)	248 (13.5)
RI	233 (3.9)	243 (10.5)	236 (3.7)	236 (3.7)
SC	266 (5.2)	249 (9.9)	261 (4.6)	261 (4.6)
SD	--	--	--	--
TN	--	--	--	--
TX	256 (2.3)	252 (10.9)	255 (2.3)	255 (2.3)
UT	226 (4.8)	229 (9.9)	231 (4.4)	230 (4.3)
VT	--	--	--	--
VA	259 (3.2)	253 (10.5)	257 (3.1)	257 (3.1)
WA	244 (3.3)	247 (7.5)	244 (3.2)	244 (3.1)
WV	--	--	--	--
WI	256 (4.9)	253 (7.1)	255 (4.4)	247 (15.5)
WY	--	--	--	--

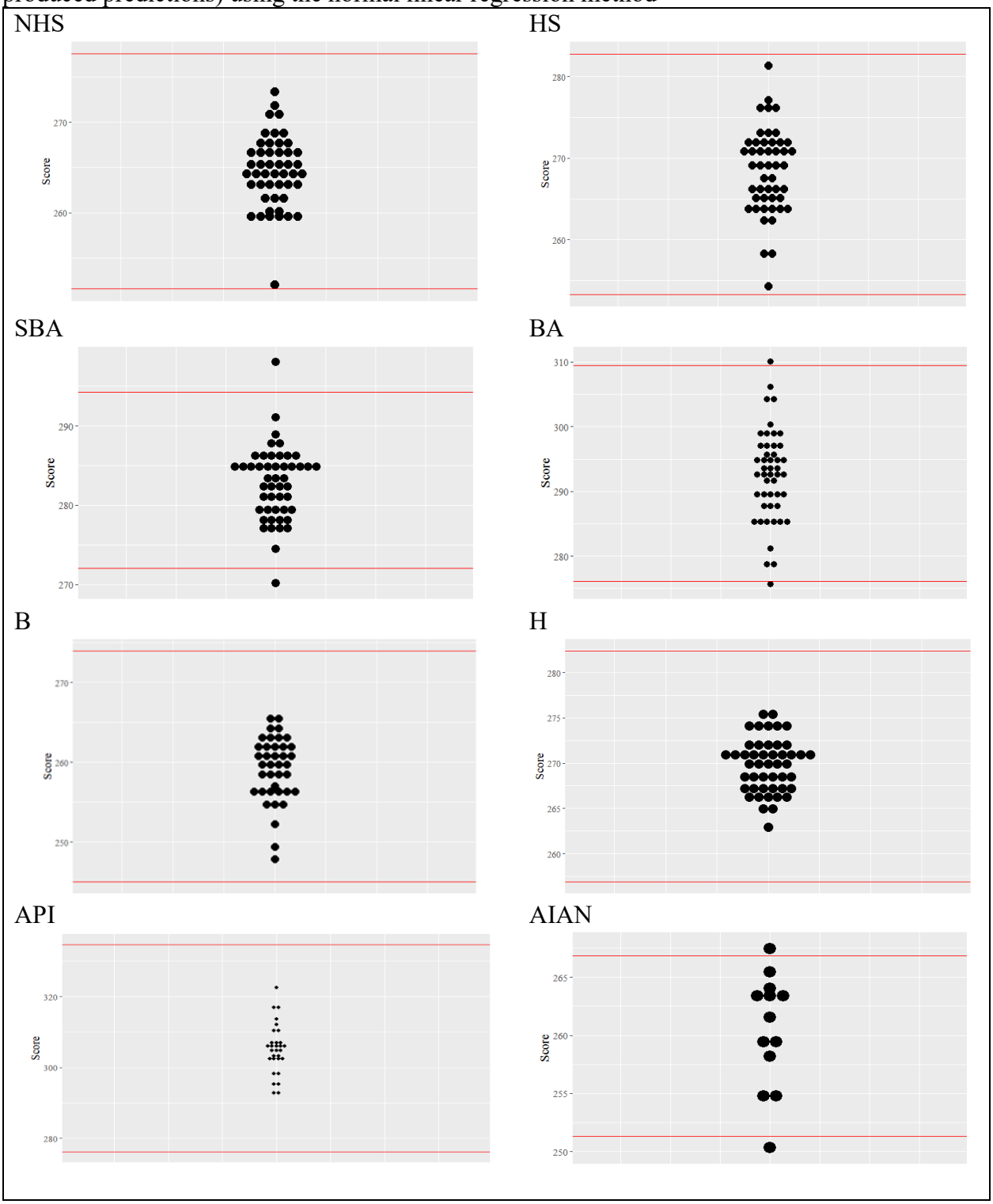
Table B.11. Supplemental table—estimates for Black students by state and technique, including NAEP-reported estimates, *with unreported NAEP estimates calculated through the FH technique*

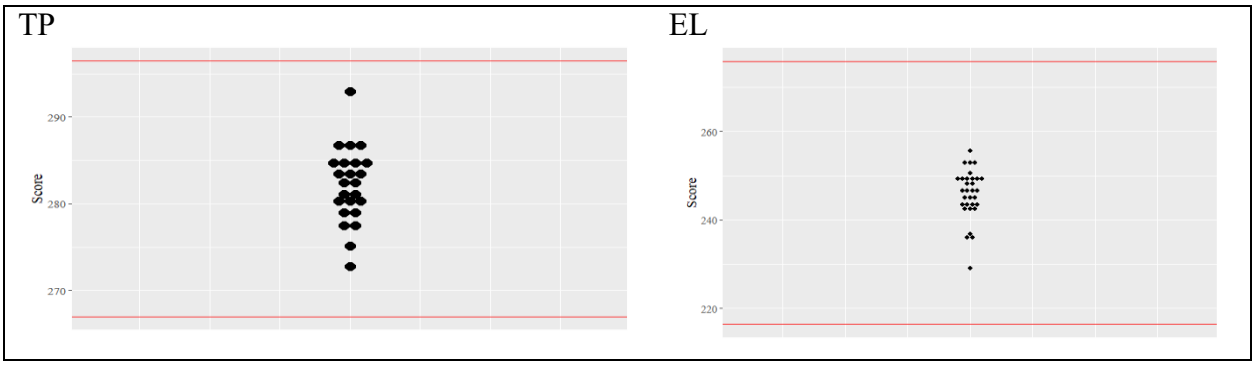
Black (B)	NAEP Reported	MICE	FH	FLEX CS
AL	248 (1.8)	249 (4.8)	254 (2.7)	256 (6.4)
AK	269 (4.5)	264 (4.5)	263 (3.7)	261 (3.8)
AZ	269 (4.1)	263 (4.7)	263 (3.5)	264 (5.5)
AR	255 (2.2)	258 (4.7)	254 (3.2)	254 (3.2)
CA	260 (3.3)	255 (4.5)	263 (4.2)	259 (7.5)
CO	260 (4.9)	261 (4.4)	265 (3.8)	265 (3.8)
CT	256 (2.9)	252 (5.4)	261 (3.5)	263 (7.3)
DE	263 (1.3)	259 (4.2)	260 (3.4)	259 (3.2)
FL	258 (2.2)	259 (2.3)	260 (4.0)	260 (4.0)
GA	264 (1.5)	260 (4.6)	258 (3.5)	262 (3.6)
HI	--	--	269 (3.3)	--
ID	--	--	261 (4.4)	--
IL	261 (2.4)	262 (4.9)	260 (2.1)	260 (2.1)
IN	257 (3.2)	263 (4.9)	257 (2.5)	258 (6.1)
IA	254 (3.0)	259 (4.8)	255 (2.5)	256 (8.5)
KS	263 (3.8)	262 (5.0)	261 (2.9)	260 (6.4)
KY	257 (2.3)	258 (5.4)	257 (2.0)	256 (2.5)
LA	255 (1.4)	256 (4.9)	255 (1.3)	255 (2.3)
ME	--	--	257 (3.6)	--
MD	263 (1.3)	260 (5.7)	263 (1.3)	264 (3.3)
MA	268 (3.6)	264 (5.7)	267 (2.9)	269 (6.5)
MI	251 (2.2)	257 (5.0)	252 (1.9)	254 (5.6)
MN	262 (2.7)	265 (5.9)	262 (2.4)	263 (3.9)
MS	257 (1.7)	257 (5.2)	257 (1.6)	257 (2.7)
MO	258 (2.7)	256 (5.5)	258 (2.2)	254 (5.7)
MT	--	--	261 (3.7)	--
NE	254 (3.6)	262 (4.9)	256 (2.7)	257 (5.8)
NV	256 (2.5)	257 (4.6)	257 (2.2)	257 (2.2)
NH	--	--	267 (3.3)	--
NJ	269 (3.0)	261 (4.6)	266 (2.5)	262 (7.3)
NM	--	--	259 (3.3)	--
NY	264 (2.9)	260 (4.5)	263 (2.5)	263 (2.4)
NC	263 (2.0)	262 (4.3)	262 (1.8)	262 (2.4)
ND	263 (4.7)	263 (4.6)	261 (3.5)	257 (6.1)
OH	259 (3.1)	257 (4.6)	258 (2.5)	258 (2.5)
OK	260 (2.7)	254 (4.5)	259 (2.3)	253 (6.7)
OR	--	--	261 (3.8)	--
PA	253 (2.3)	256 (5.2)	254 (2.0)	254 (5.2)
RI	258 (3.1)	255 (5.2)	259 (2.6)	259 (2.6)
SC	256 (1.9)	261 (5.0)	257 (1.7)	257 (2.7)
SD	--	--	260 (3.7)	--
TN	253 (3.0)	260 (4.0)	255 (2.4)	258 (4.9)
TX	267 (2.9)	266 (4.3)	265 (2.4)	264 (2.4)
UT	--	--	262 (4.1)	--
VT	--	--	263 (3.6)	--
VA	265 (1.8)	264 (5.2)	265 (1.7)	265 (1.7)
WA	257 (3.4)	260 (4.8)	260 (2.6)	260 (2.6)
WV	256 (3.1)	248 (5.6)	254 (2.7)	254 (5.2)
WI	249 (4.2)	261 (4.0)	253 (3.0)	253 (2.6)
WY	--	--	258 (4.0)	--

Appendix C: Supplemental Plots

Plots for Evaluating Plausibility of Preliminary Sets of Mice Imputations

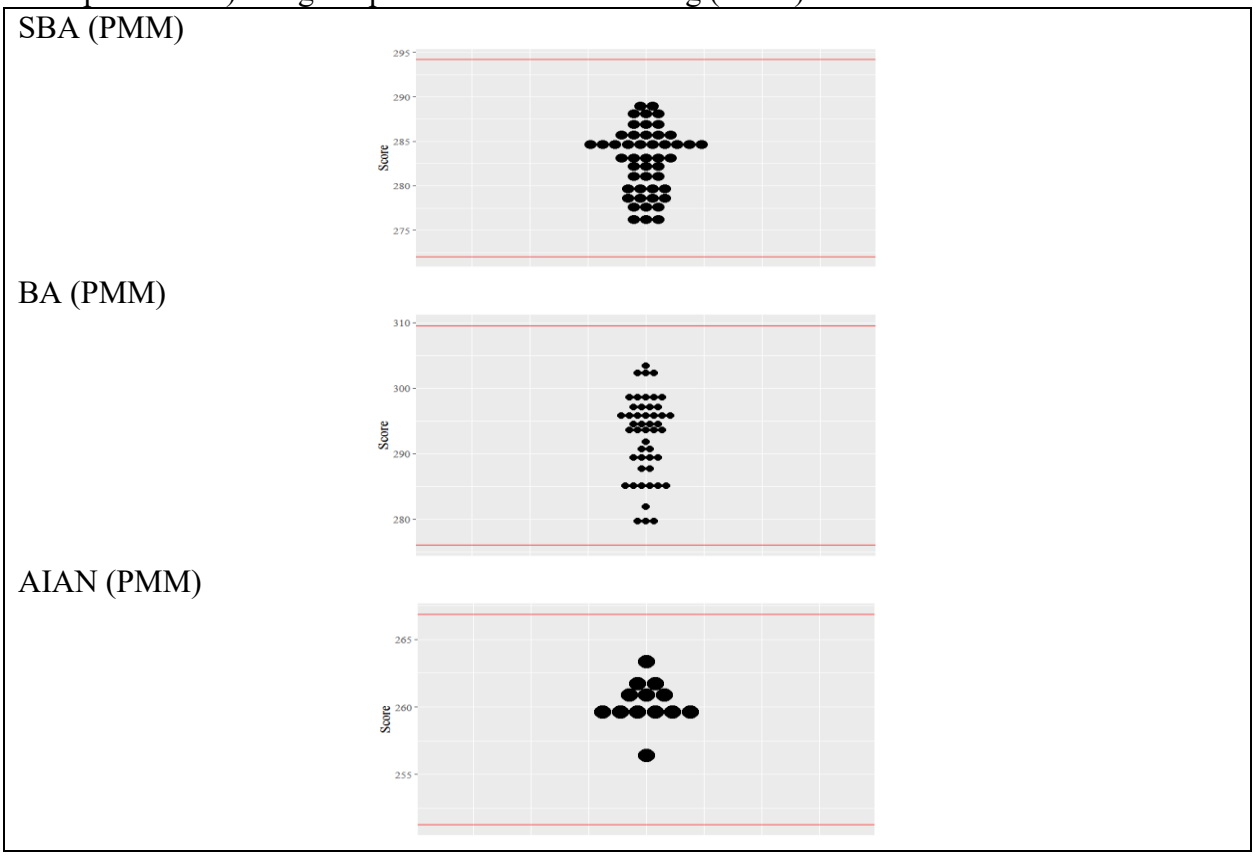
Figure C.1.1. Plots for evaluating plausibility of preliminary sets of mice imputations (MICE-produced predictions) using the normal linear regression method





Two predicted values are out of bounds for the SBA subgroup, AL is lower and MA is higher. Two predicted values are also out of bounds for the BA subgroup. Again, AL is lower and MA is higher. Two predicted values are out of bounds for the AIAN subgroup. AZ is lower and MN is higher. The procedure is re-run for these subgroups, but with PMM instead of normal linear regression.

Figure C.1.2. Plots for evaluating plausibility of preliminary sets of mice imputations (MICE-based predictions) using the predictive mean matching (PMM) method



Coverage Plots: Target Intervals and Technique-produced Estimates by State

MICE

Figure C.1.3. MICE-produced estimates and target intervals for NHS subgroup by state

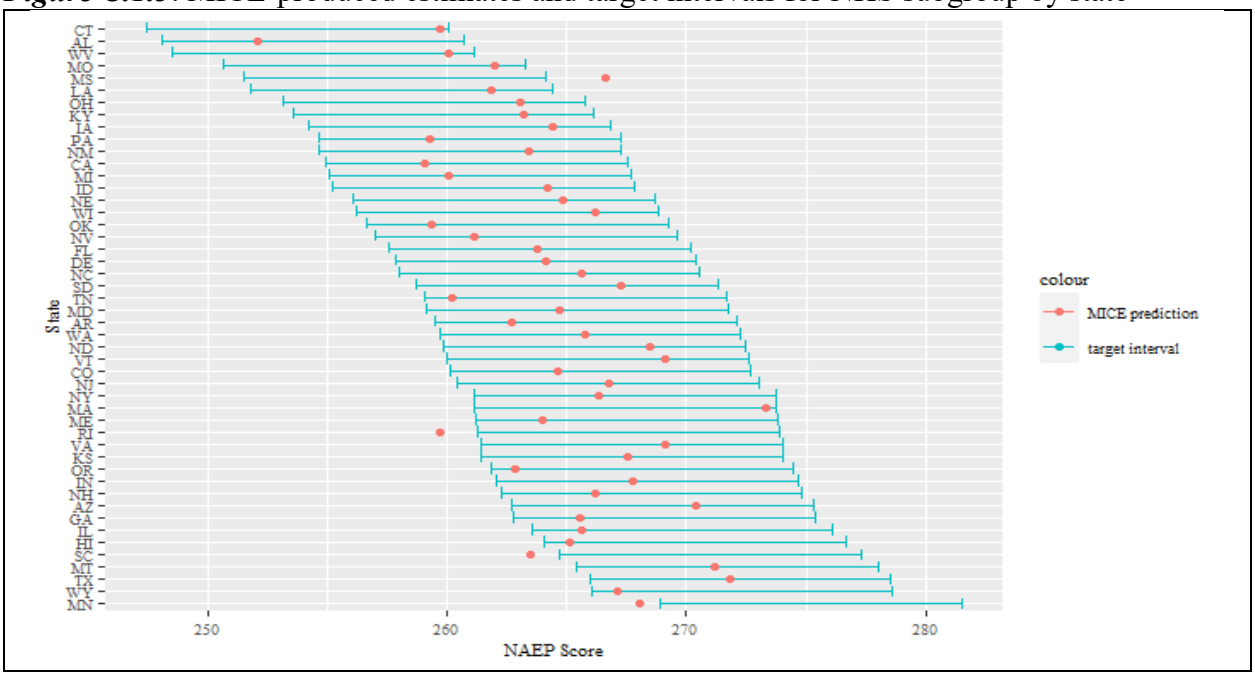


Figure C.1.4. MICE-produced estimates and target intervals for HS subgroup by state

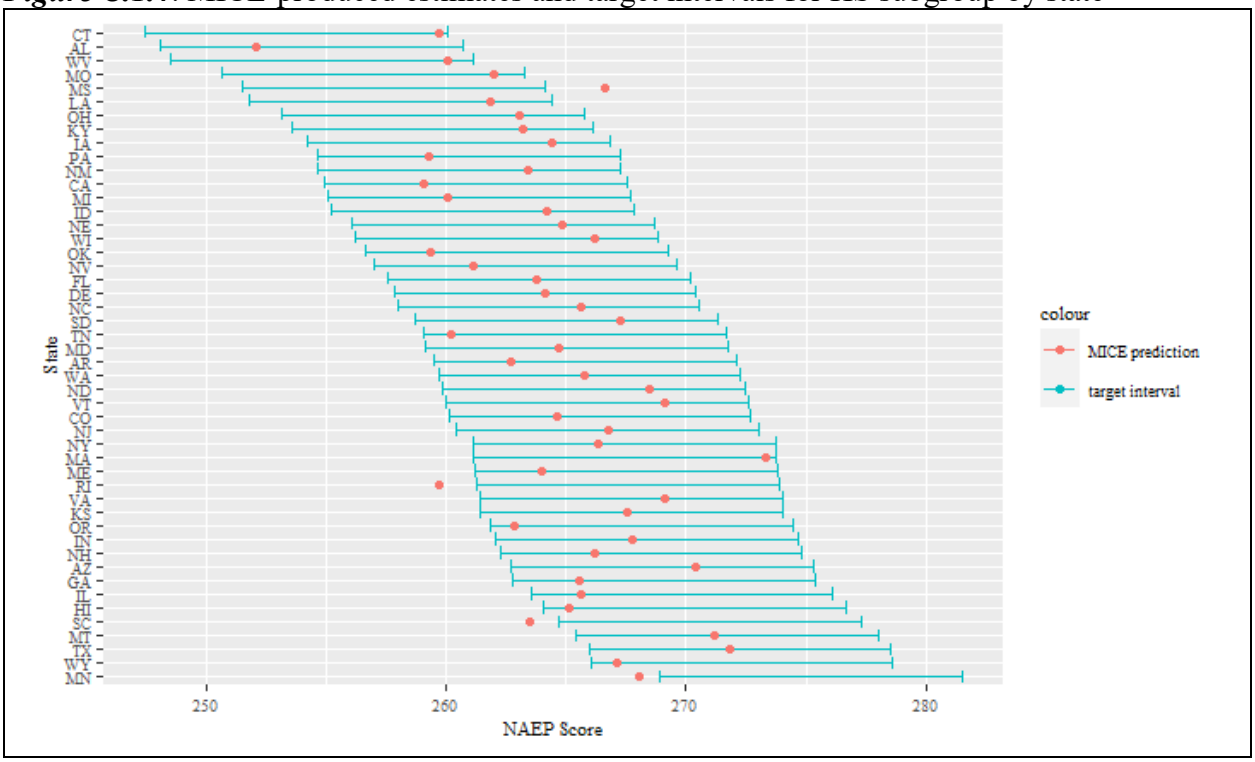


Figure C.1.5. MICE-produced estimates and target intervals for SBA subgroup by state

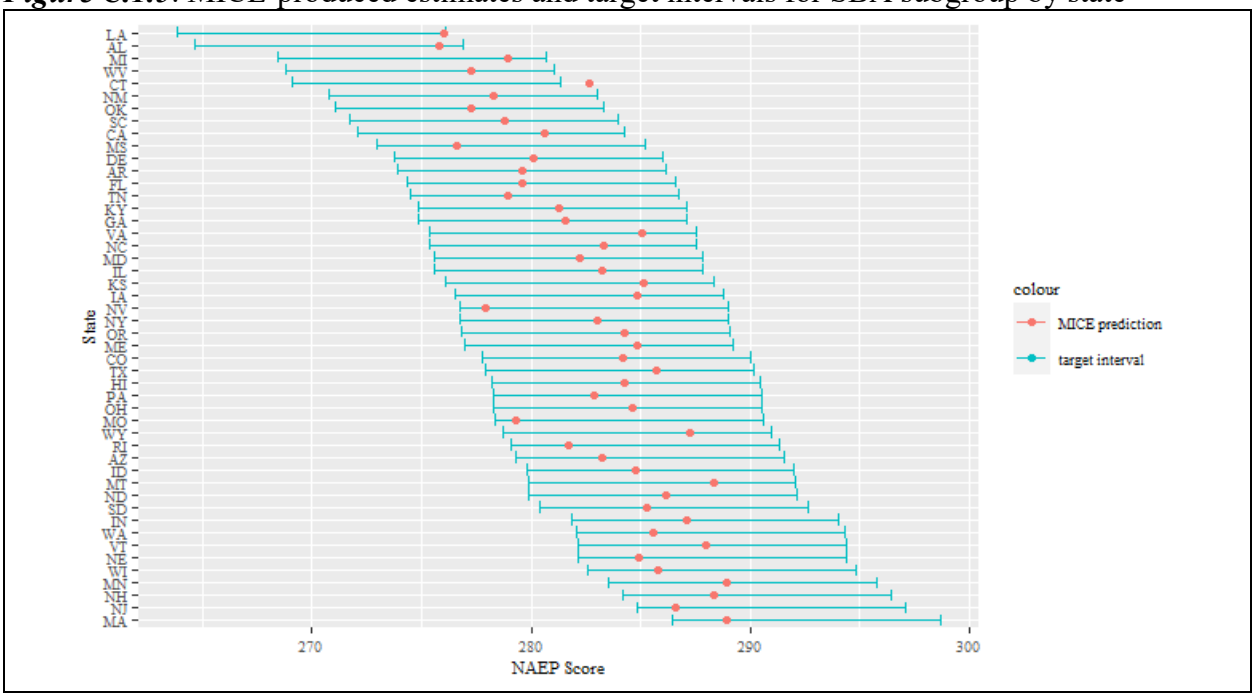


Figure C.1.6. MICE-produced estimates and target intervals for BA subgroup by state

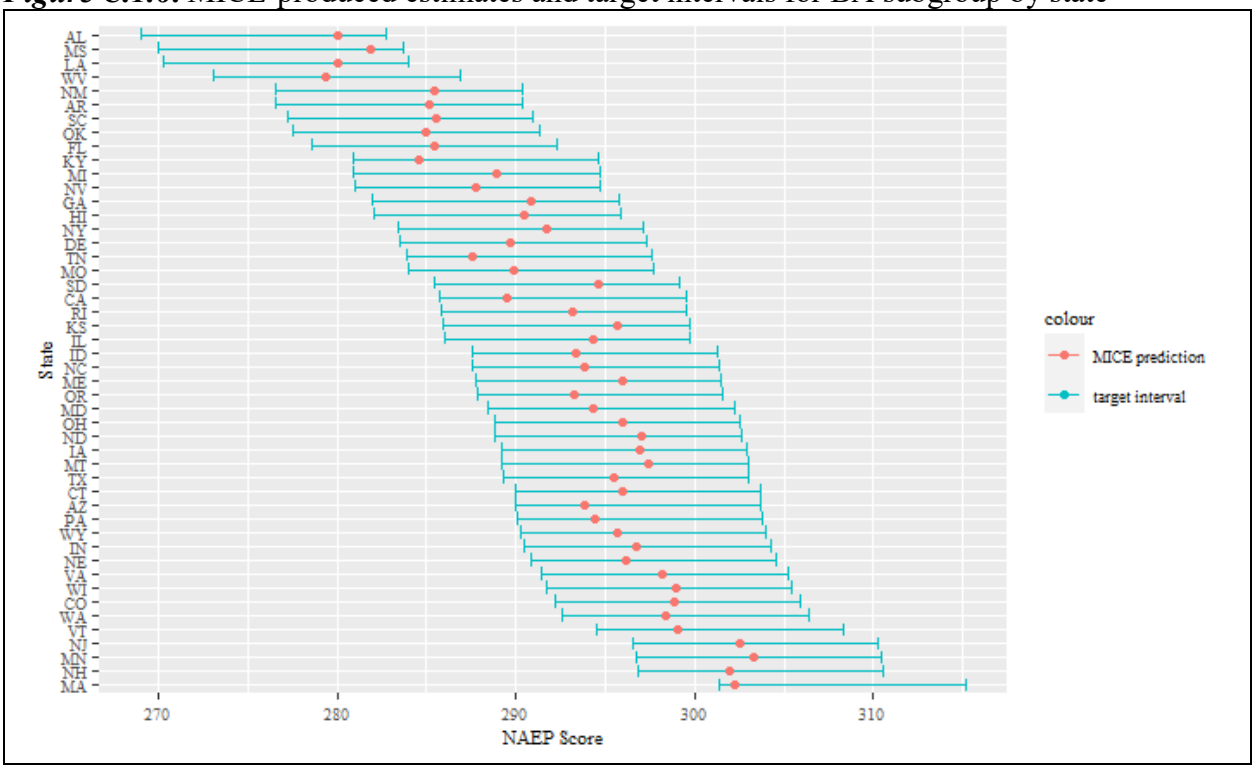


Figure C.1.7. MICE-produced estimates and target intervals for B subgroup by state

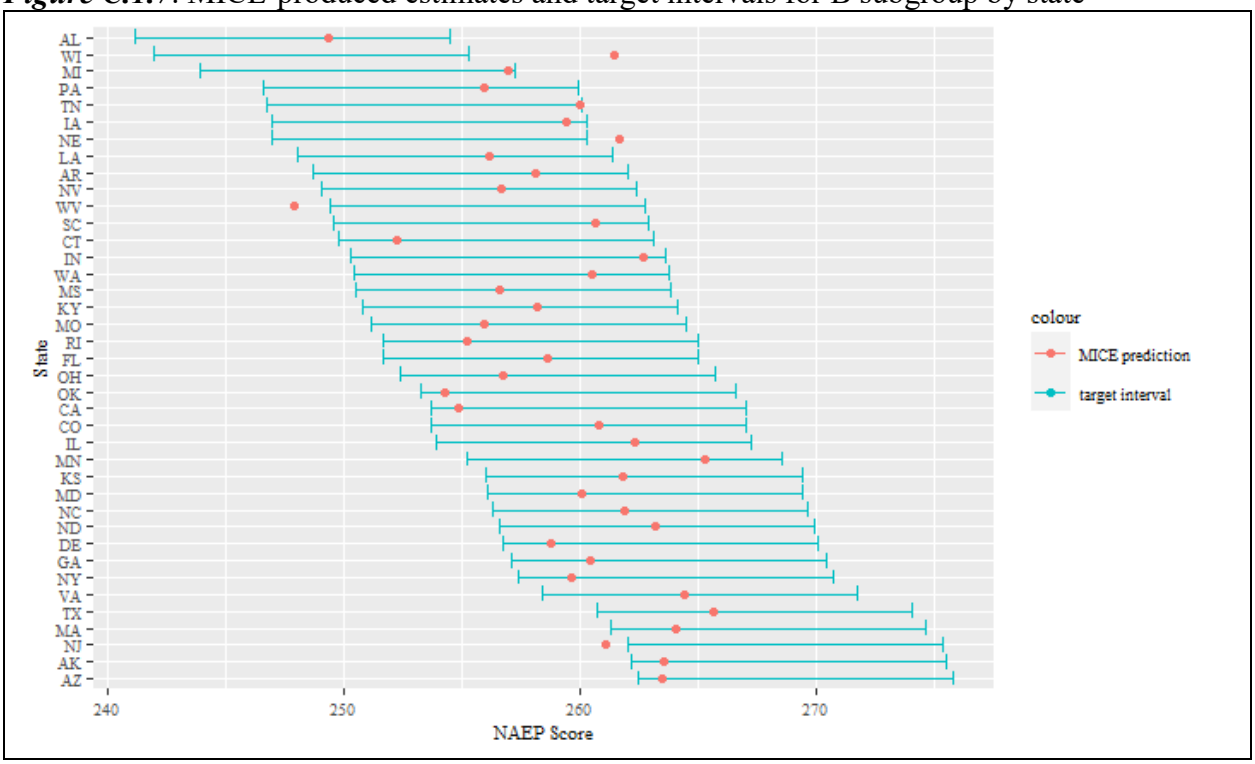


Figure C.1.8. MICE-produced estimates and target intervals for H subgroup by state

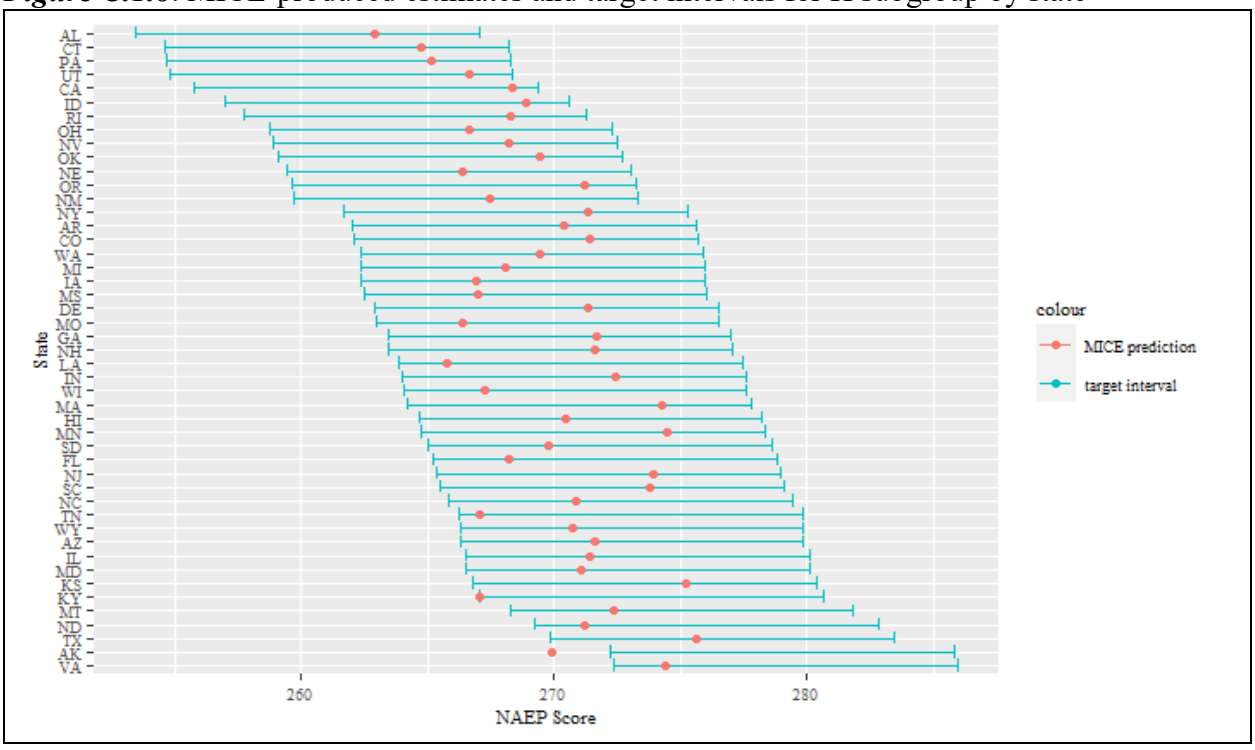


Figure C.1.9. MICE-produced estimates and target intervals for API subgroup by state

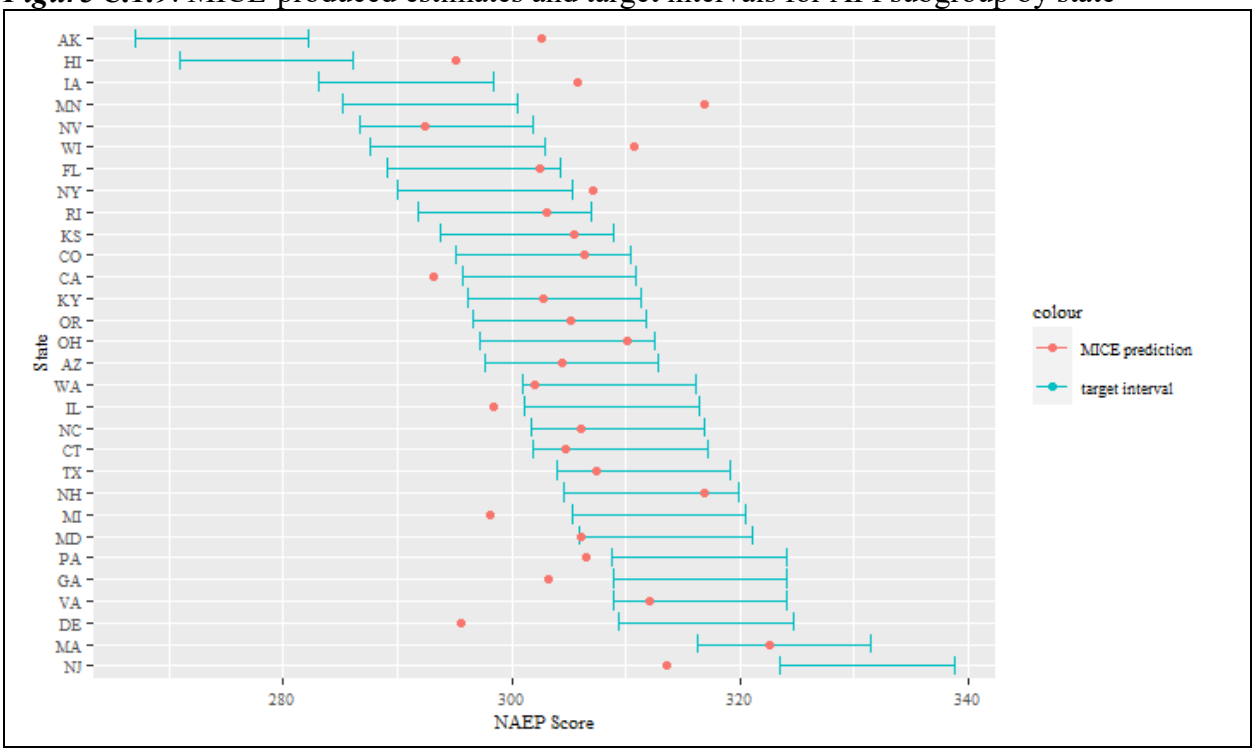


Figure C.1.10. MICE-produced estimates and target intervals for AIAN subgroup by state

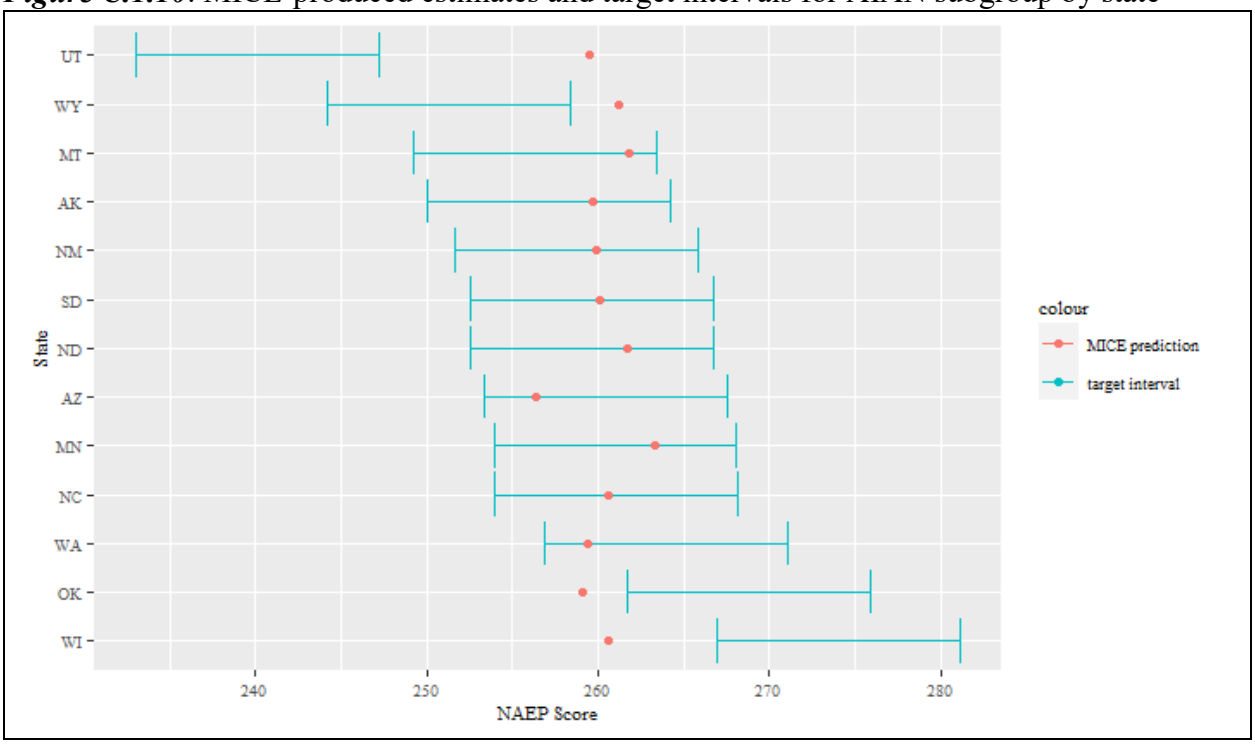


Figure C.1.11. MICE-produced estimates and target intervals for TP subgroup by state

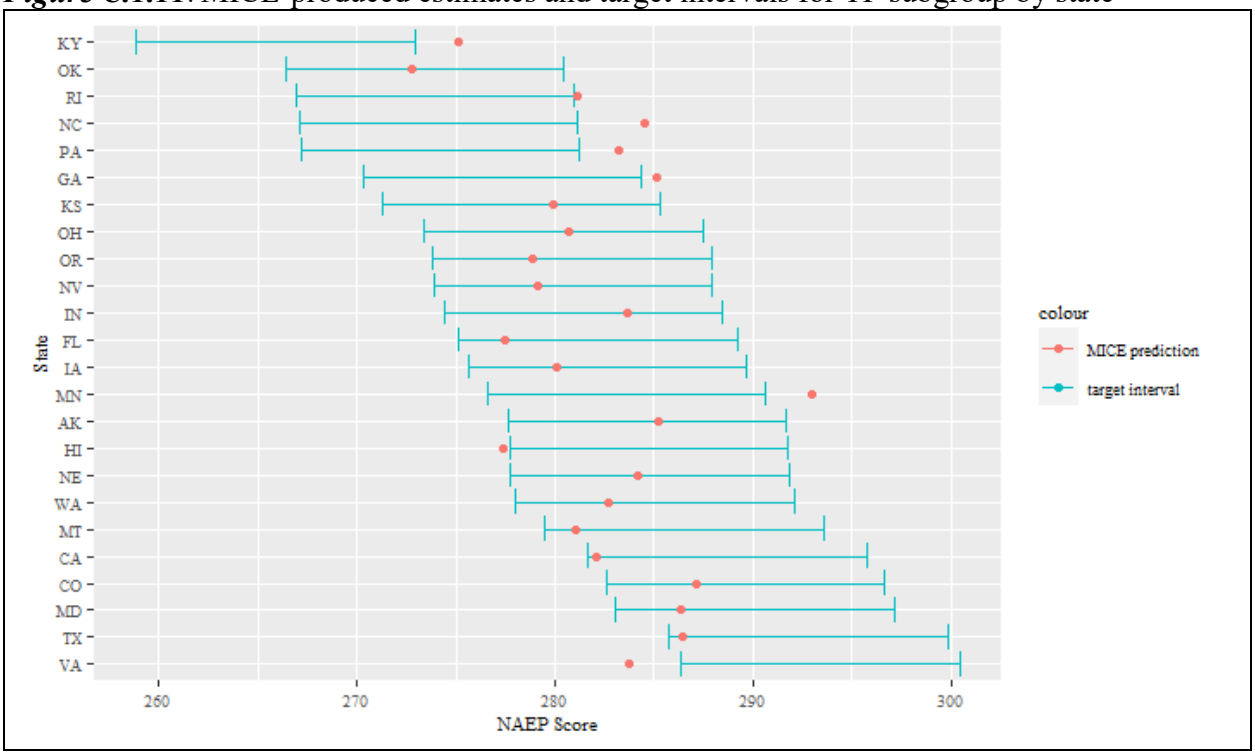
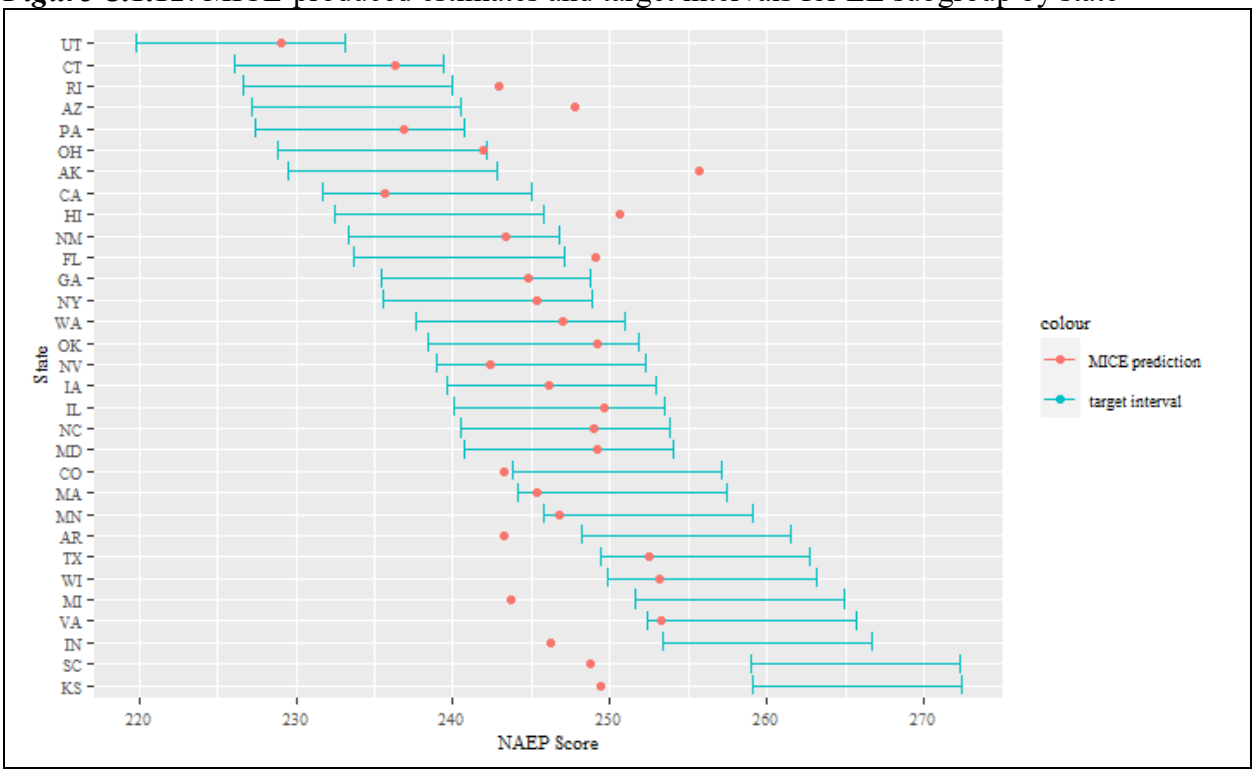


Figure C.1.12. MICE-produced estimates and target intervals for EL subgroup by state



FH

Figure C.2.1. FH-produced estimates and target intervals for NHS subgroup by state

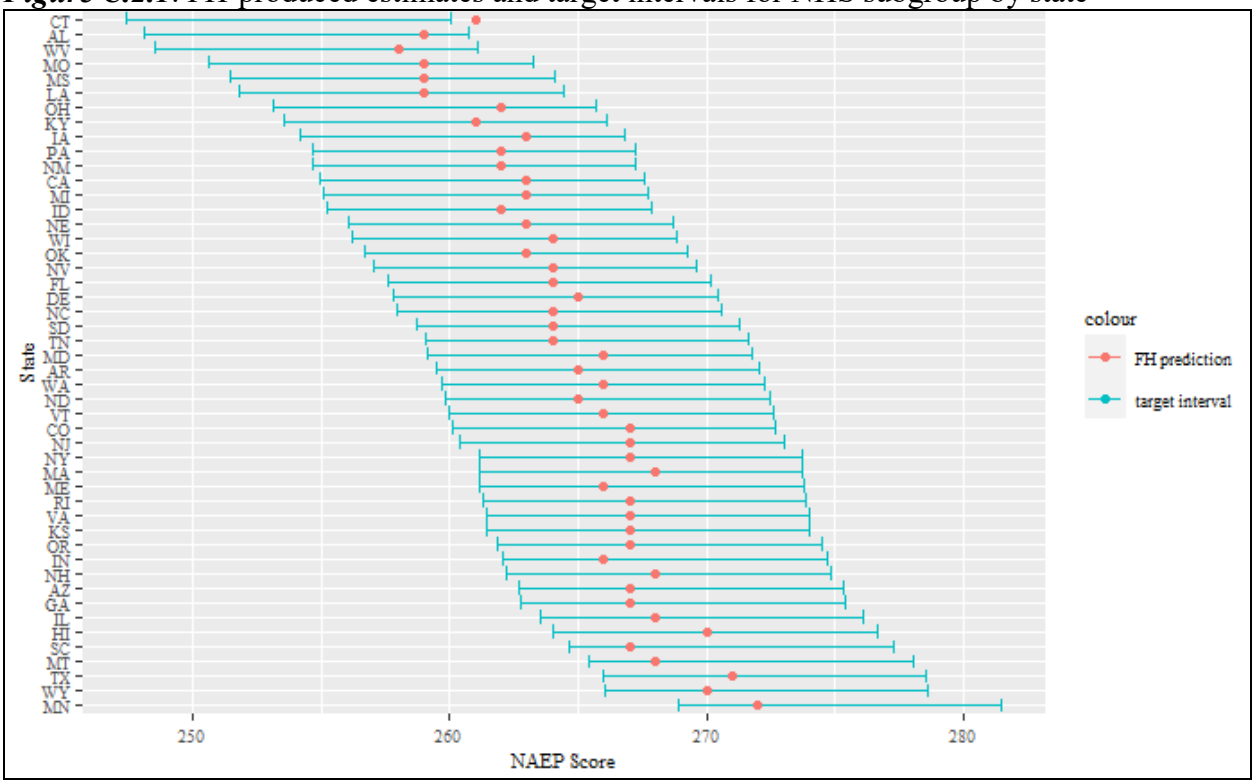


Figure C.2.2. FH-produced estimates and target intervals for HS subgroup by state

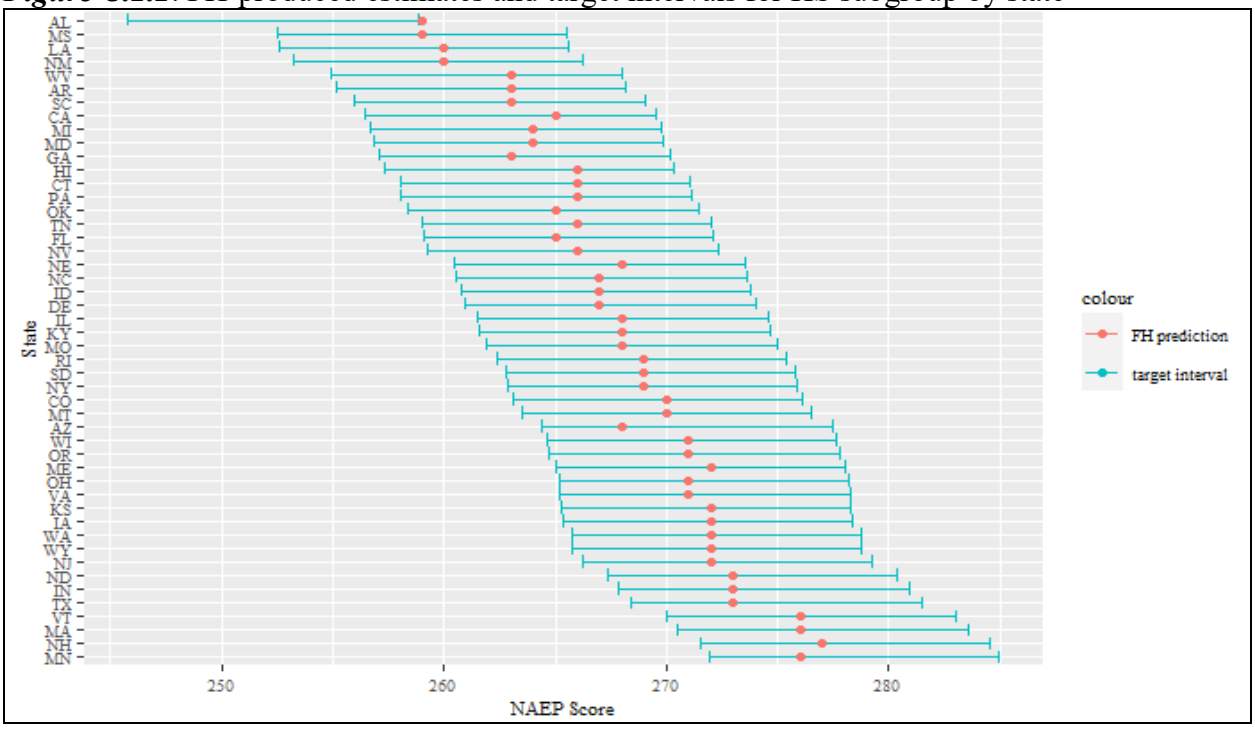


Figure C.2.3. FH-produced estimates and target intervals for SBA subgroup by state

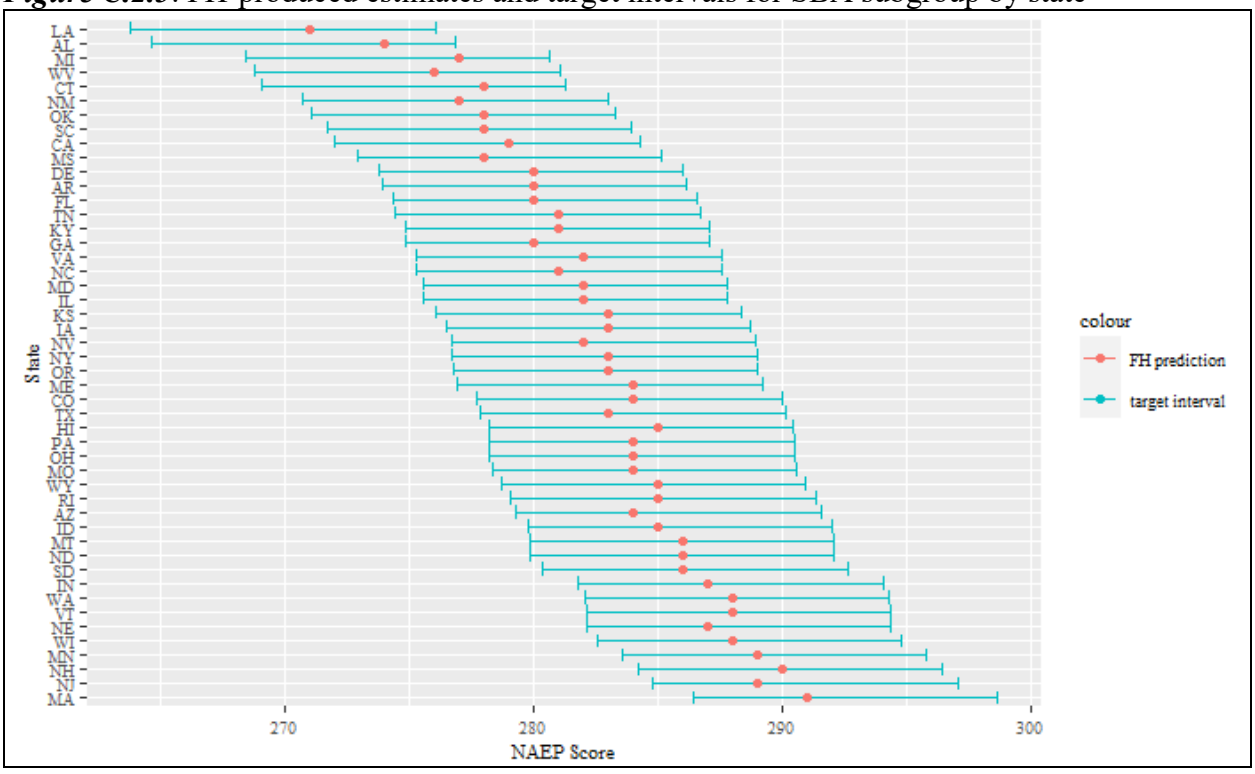


Figure C.2.4. FH-produced estimates and target intervals for BA subgroup by state

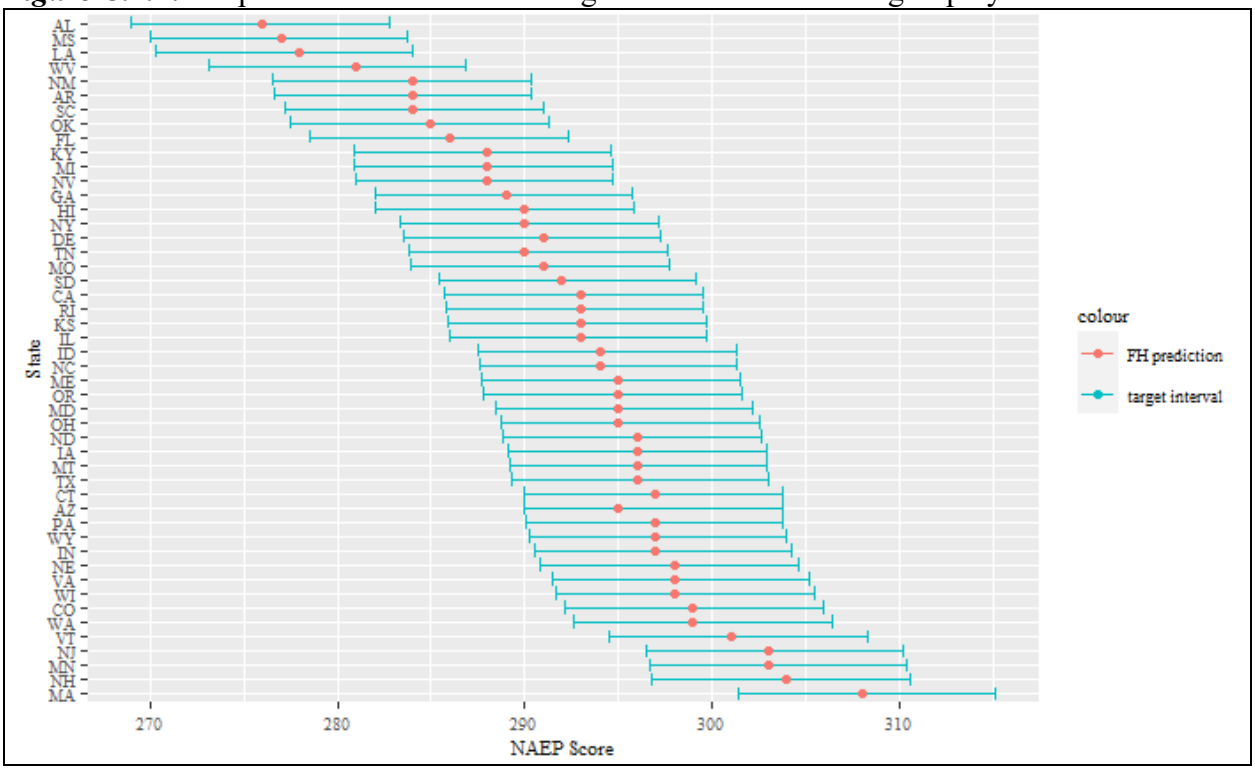


Figure C.2.5. FH-produced estimates and target intervals for B subgroup by state

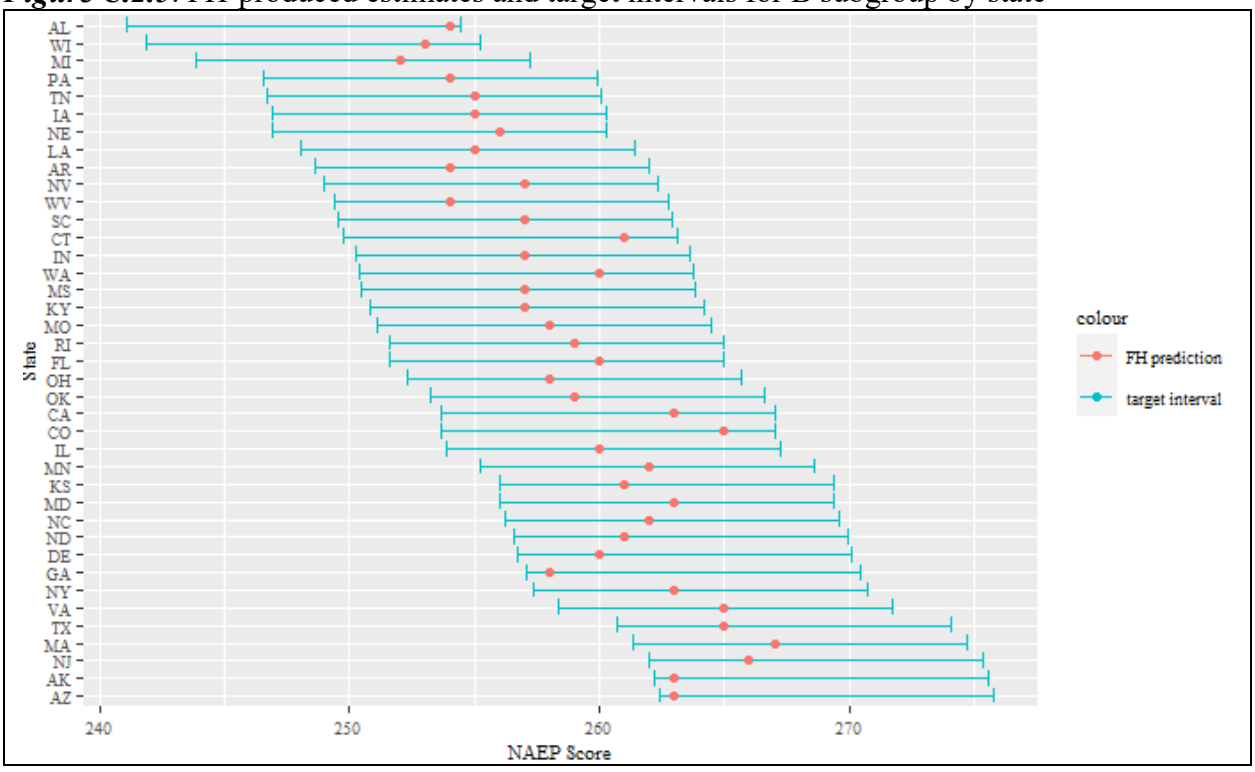


Figure C.2.6. FH-produced estimates and target intervals for H subgroup by state

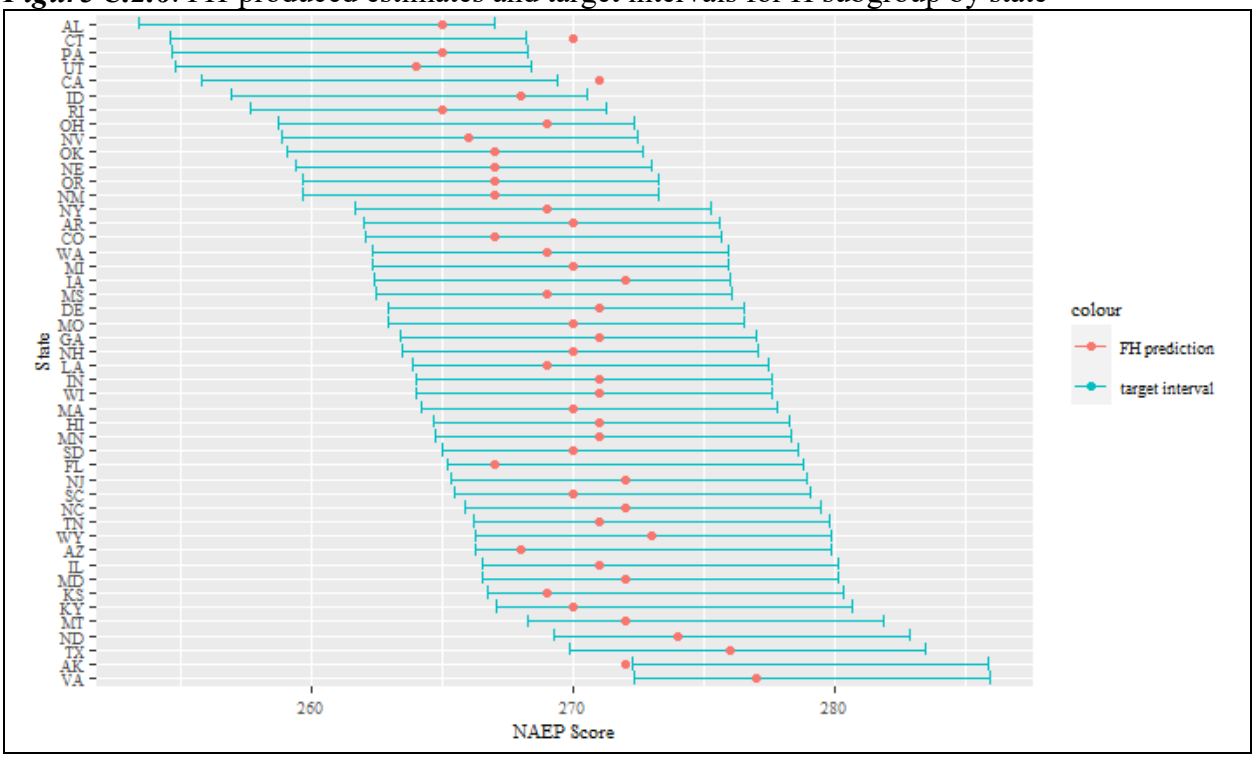


Figure C.2.7. FH-produced estimates and target intervals for API subgroup by state

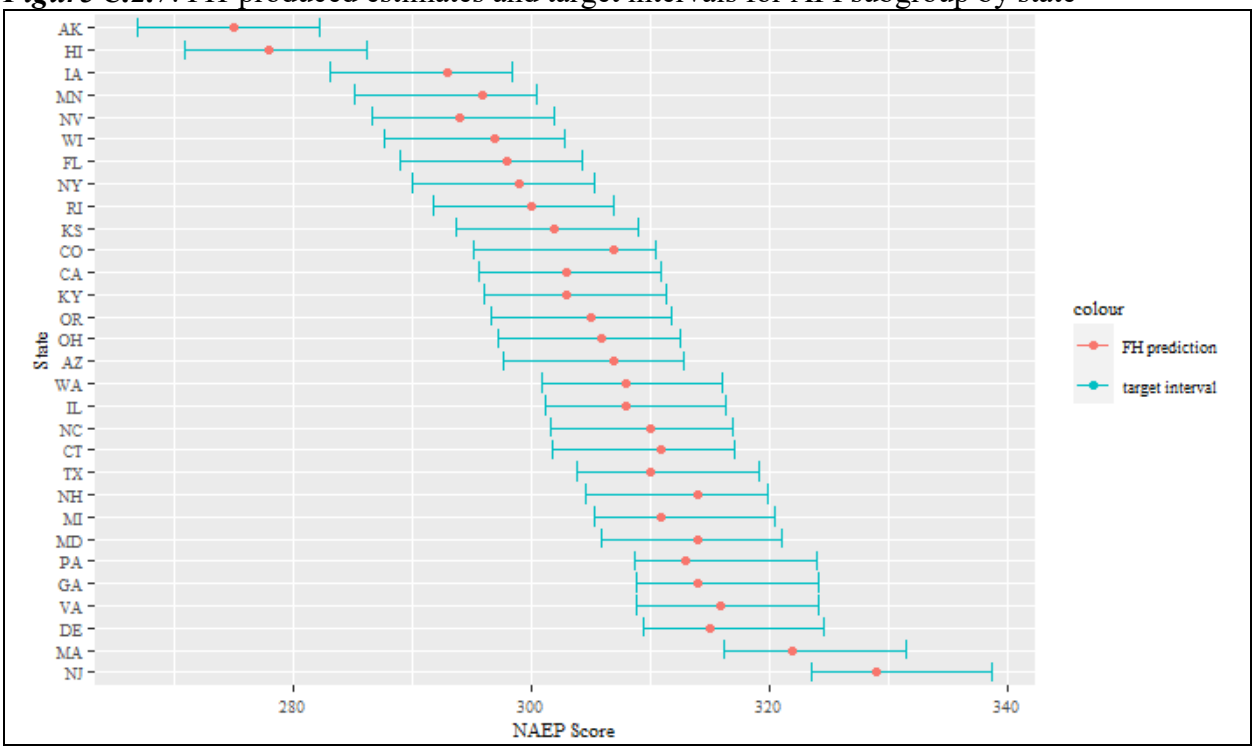


Figure C.2.8. FH-produced estimates and target intervals for AIAN subgroup by state

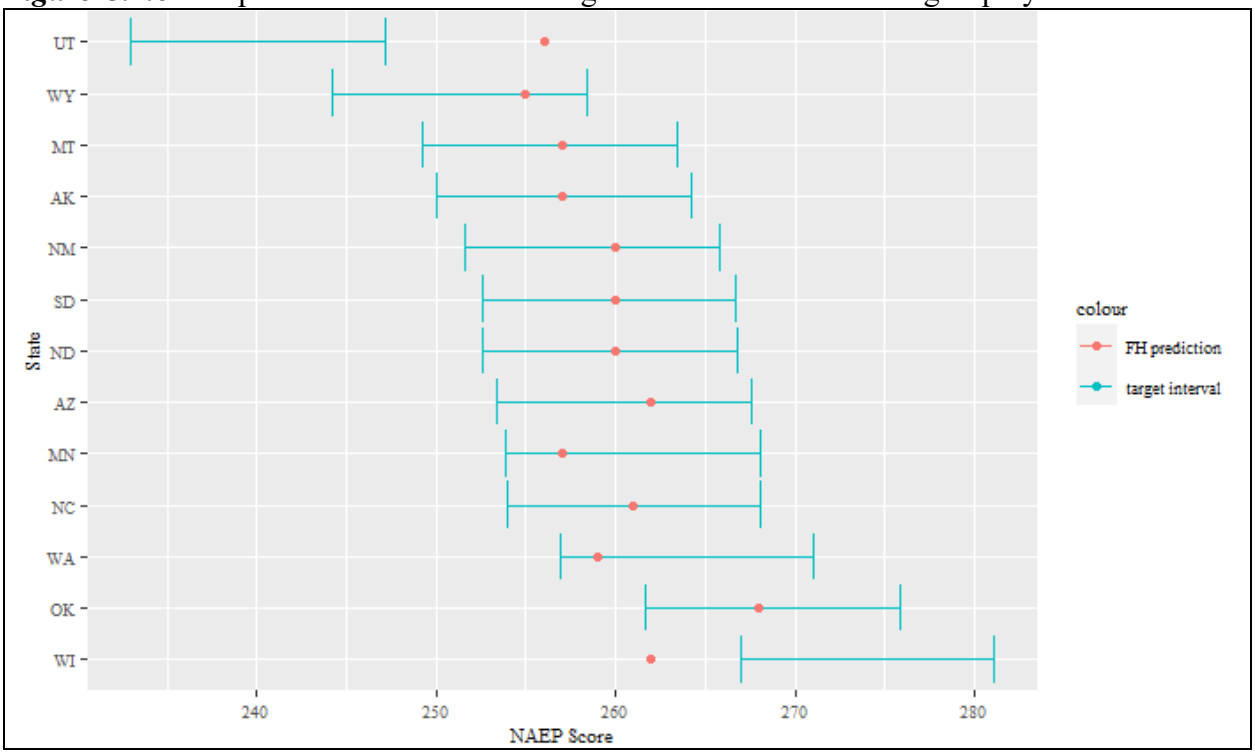


Figure C.2.9. FH-produced estimates and target intervals for TP subgroup by state

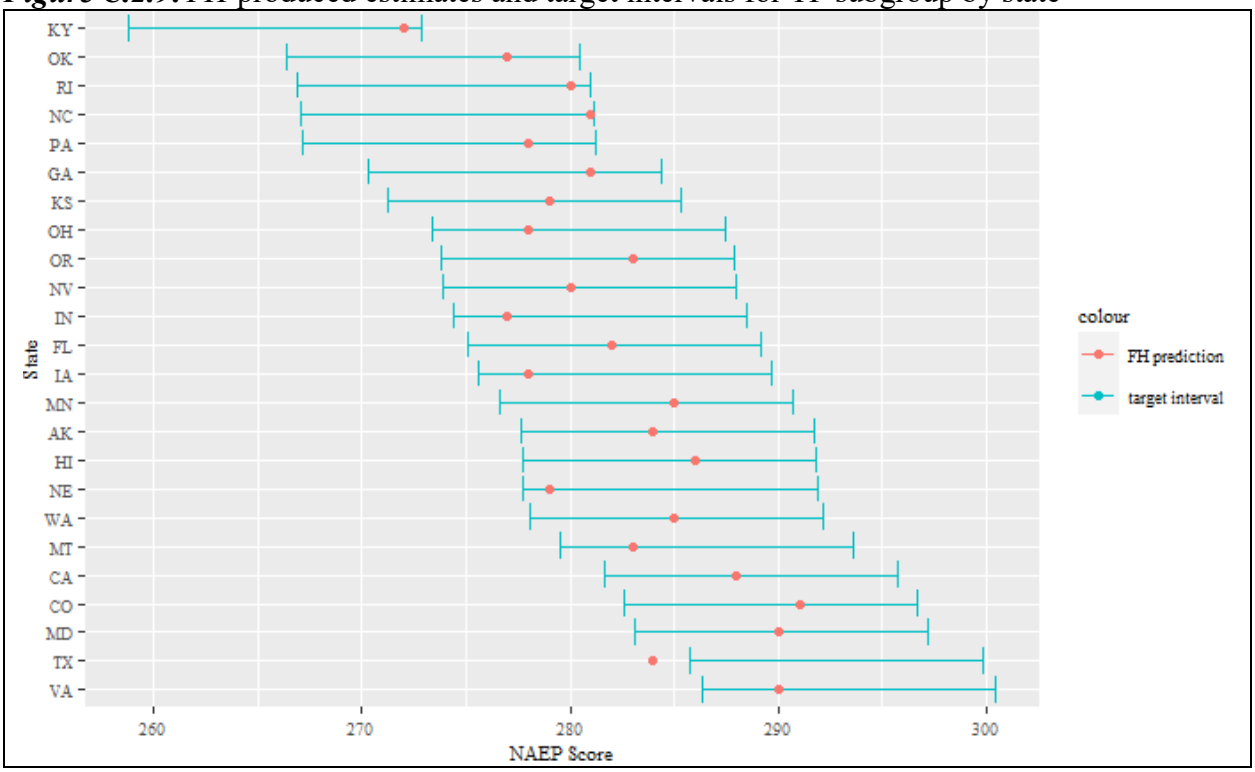
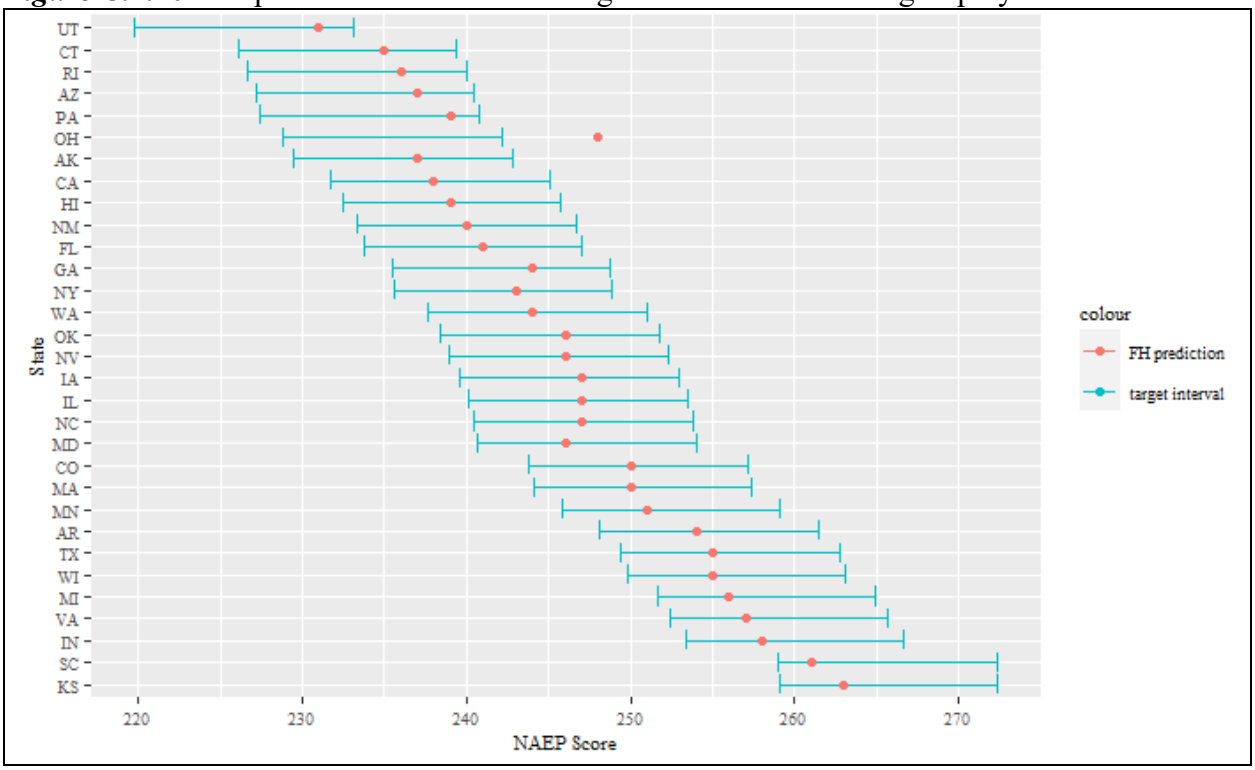


Figure C.2.10. FH-produced estimates and target intervals for EL subgroup by state



FLEX CS

Figure C.3.1. FLEX CS-produced estimates and target intervals for NHS subgroup by state

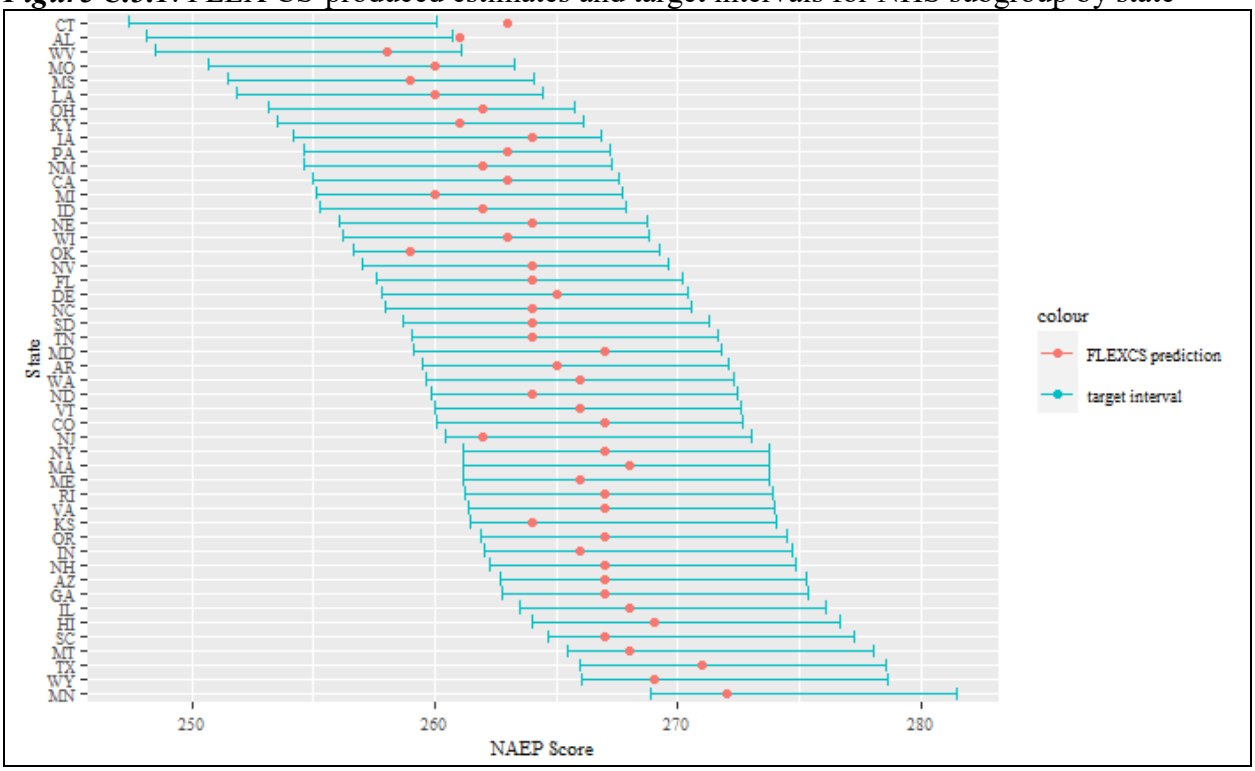


Figure C.3.2. FLEX CS-produced estimates and target intervals for HS subgroup by state

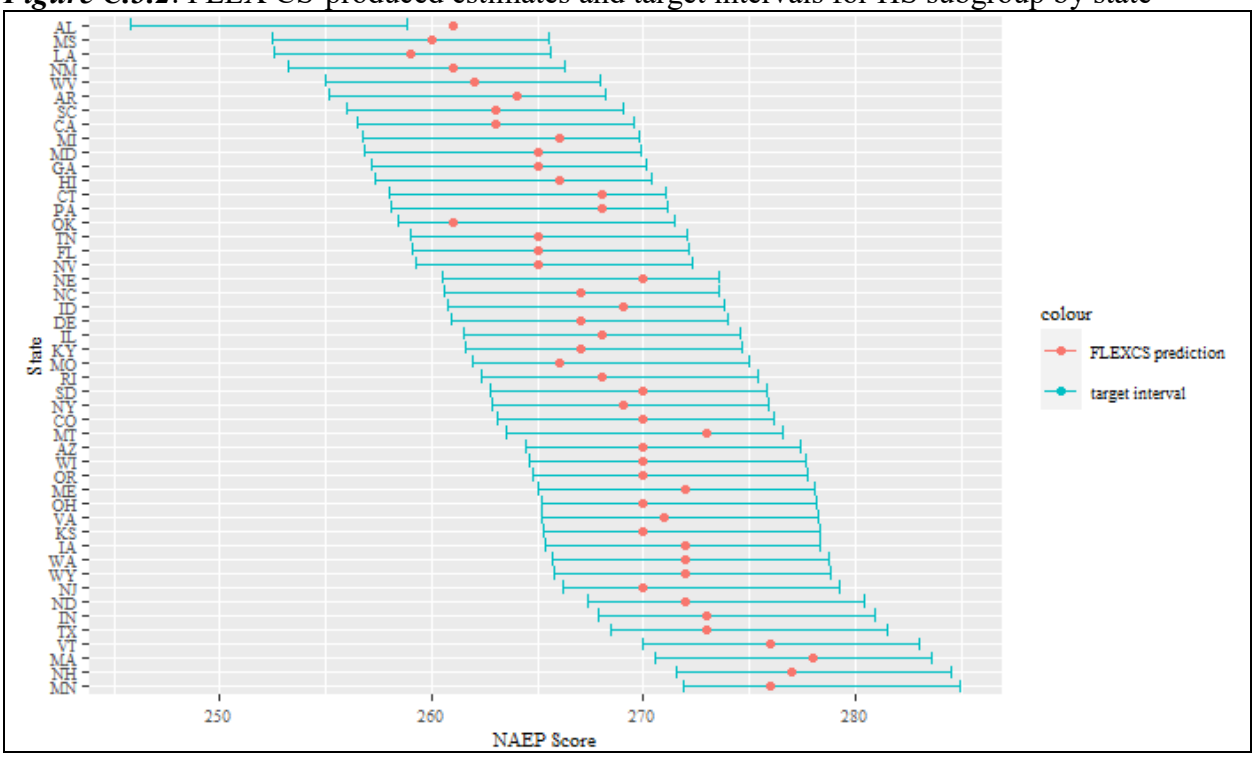


Figure C.3.3. FLEX CS-produced estimates and target intervals for SBA subgroup by state

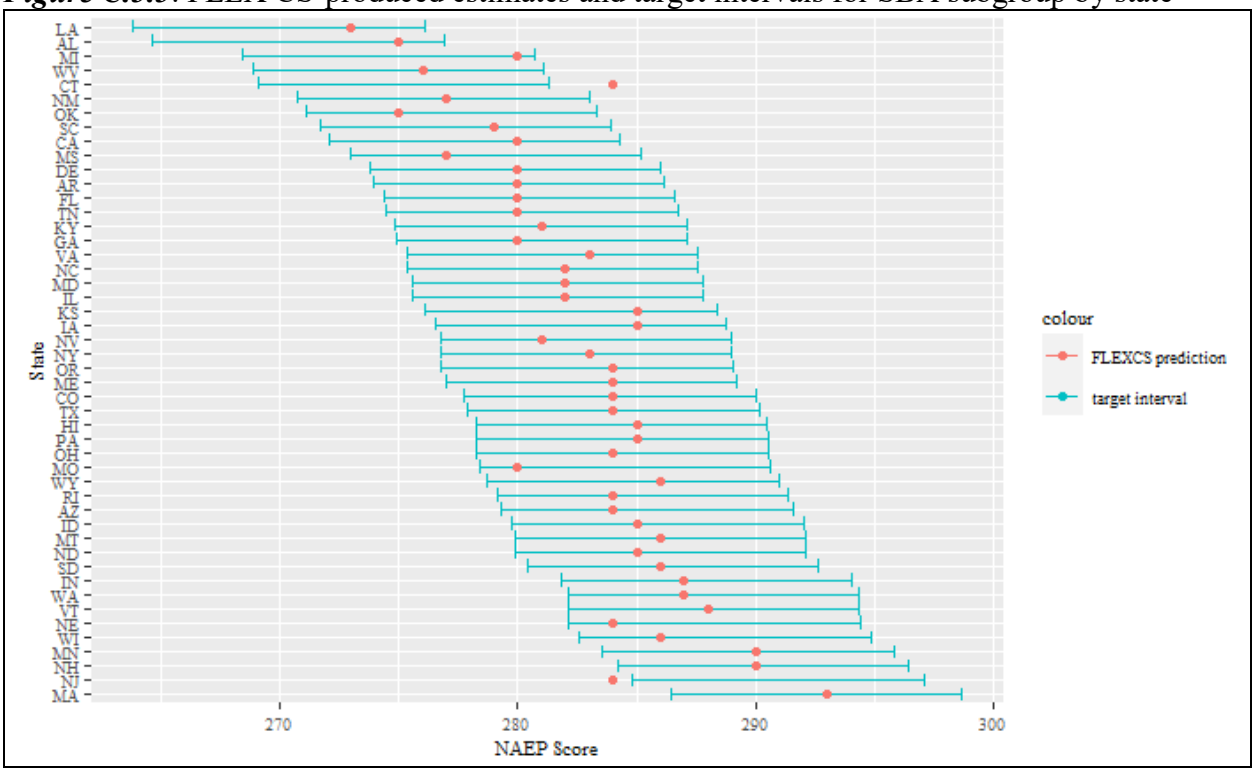


Figure C.3.4. FLEX CS-produced estimates and target intervals for BA subgroup by state

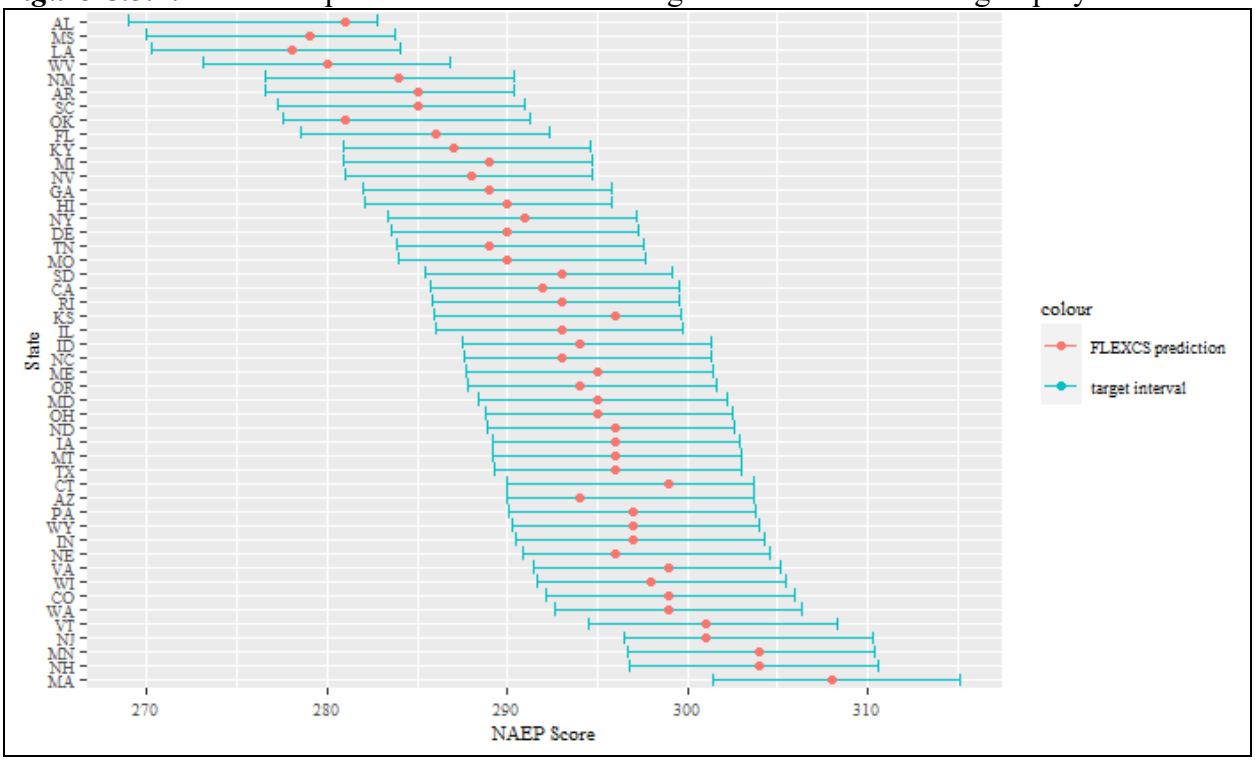


Figure C.3.5. FLEX CS-produced estimates and target intervals for B subgroup by state

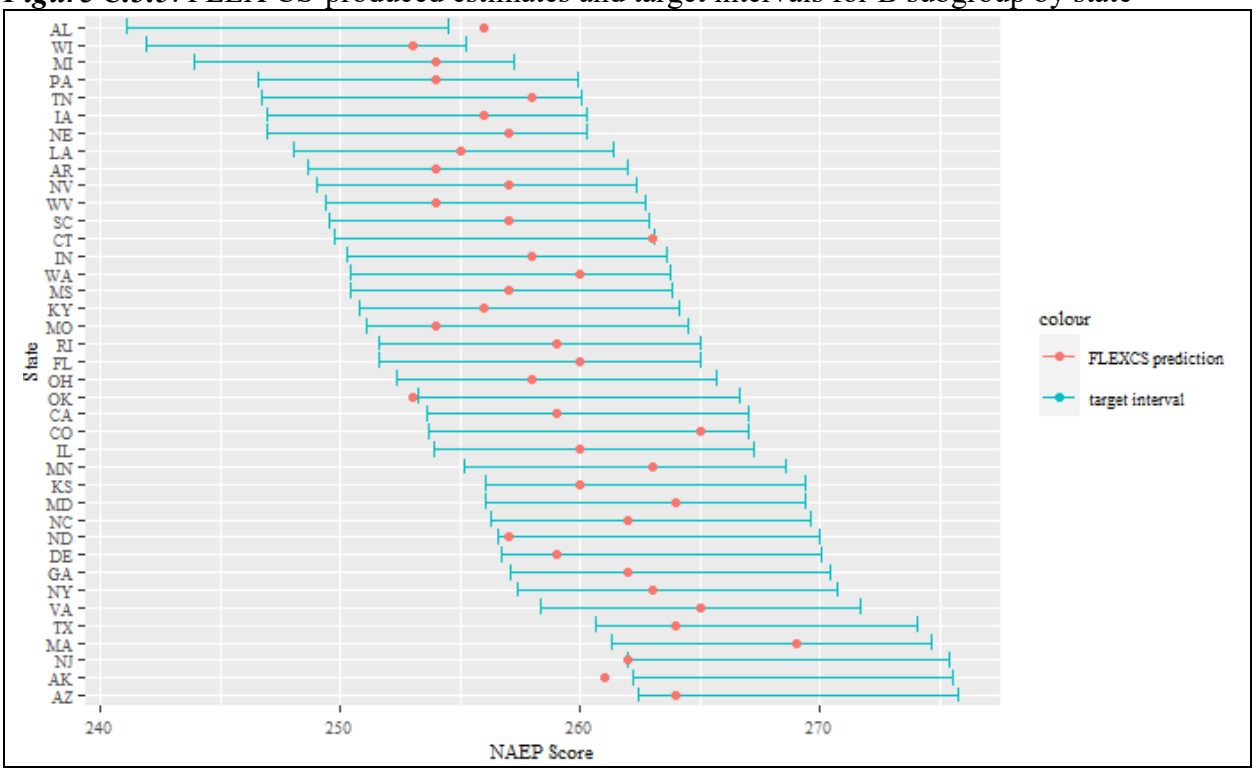


Figure C.3.6. FLEX CS-produced estimates and target intervals for H subgroup by state

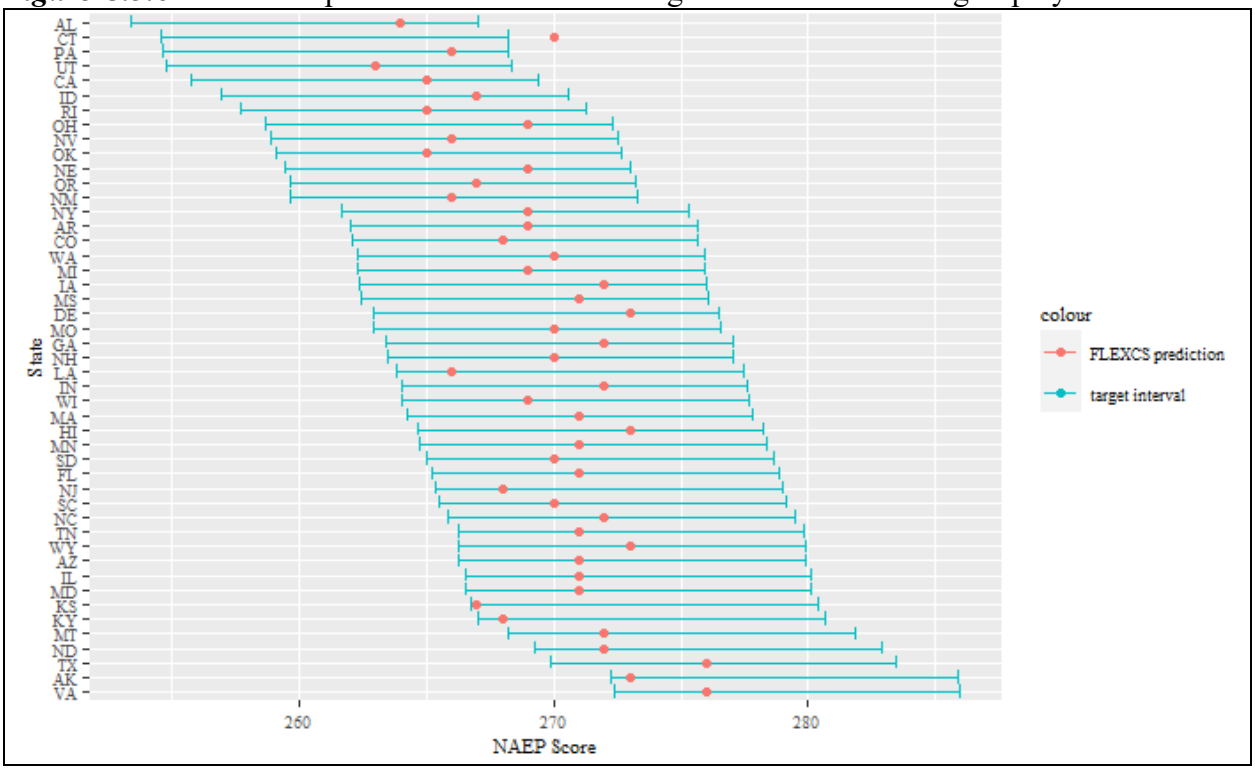


Figure C.3.7. FLEX CS-produced estimates and target intervals for API subgroup by state

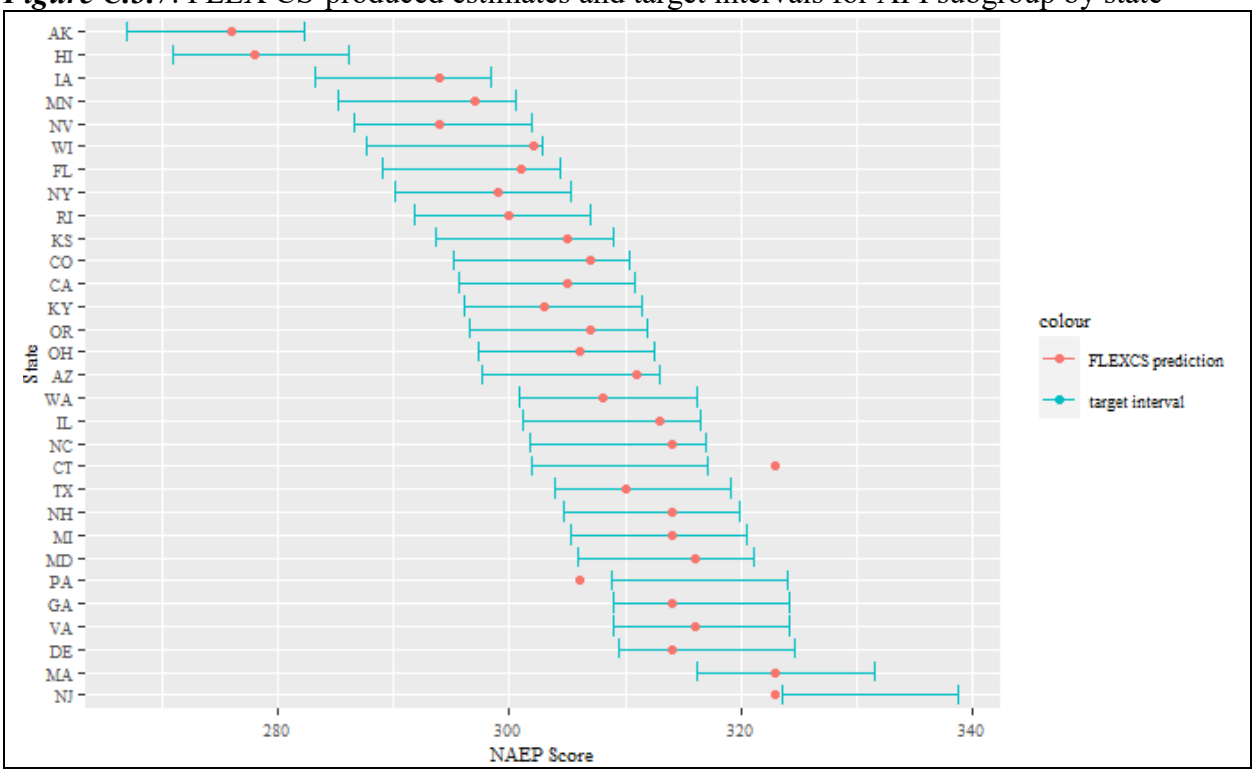


Figure C.3.8. FLEX CS-produced estimates and target intervals for AIAN subgroup by state

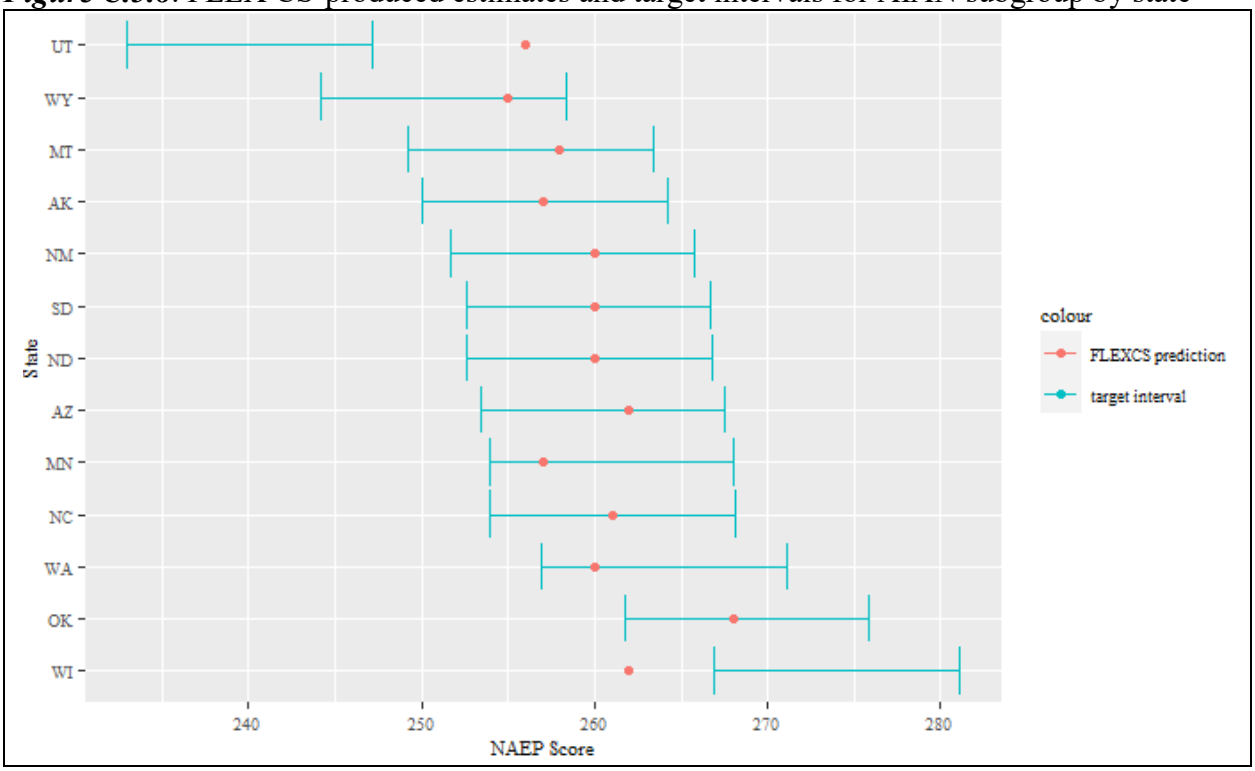


Figure C.3.9. FLEX CS-produced estimates and target intervals for TP subgroup by state

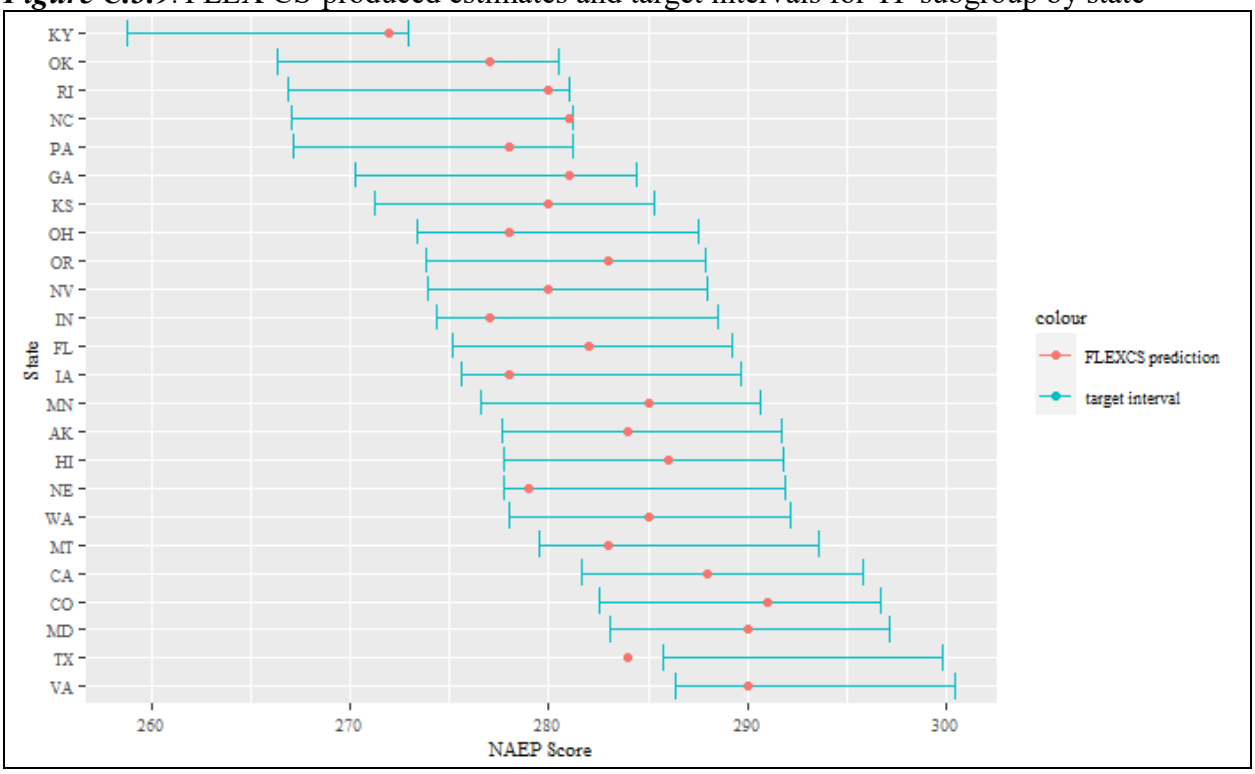


Figure C.3.10. FLEX CS-produced estimates and target intervals for EL subgroup by state