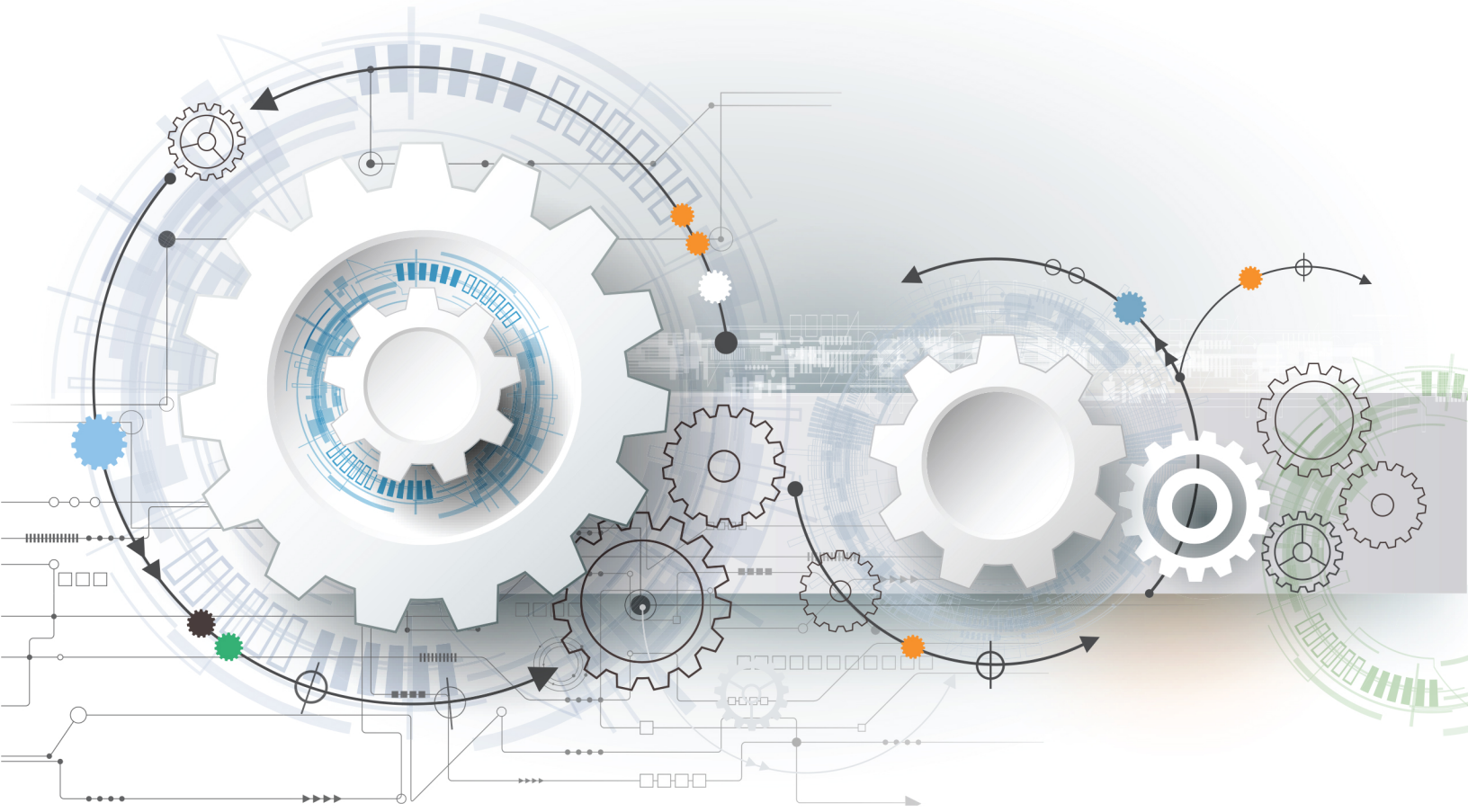
Working Paper

2018-3

Estimation of Nationally-Referenced Status and Growth Norms Using Multiple Imputation

JEFF ALLEN, PHD

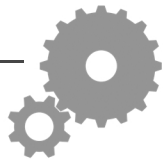


ACT.org

ABOUT THE AUTHORS

Jeff Allen is a statistician in the Research division at ACT. He specializes in longitudinal research linking test scores to educational outcomes and student growth models.

ACT WORKING PAPER SERIES



ACT working papers document preliminary research. The papers are intended to promote discussion and feedback before formal publication. The research does not necessarily reflect the views of ACT.

R1701

Abstract

This study evaluates an imputation-based procedure for estimating nationally-referenced status and growth norms. Student, school, and district variables observed for both the sample and population are the basis for imputation. The procedure is applied to data from an assessment system that spans grades 3-8 to produce estimates of nationally-referenced status and conditional status norms. The estimated average population scores are higher than the sample scores in mathematics, but very similar in reading.

The performance of the estimation method is evaluated using a holdout method for both system missingness (due to system-wide nonparticipation in the assessment system) and intermittent missingness (due to missing data in one or more grade levels). For system missingness, the average absolute difference of imputed mean and actual mean (in standard deviation units) was 0.07. In 79% of the holdout comparisons, the imputed mean was within 0.10 standard deviations of the actual mean. The imputed standard deviation was within 10% of the actual standard deviation in 96% of the holdout comparisons. In 85% of cases, percentile ranks estimated from imputed data were within five points of those estimated from the actual data; 56% of cases were within two points. In 92% of cases, the residual rank (similar to student growth percentile) estimated from imputed data was within five points of that estimated with actual data. The estimation method was very accurate for intermittent missingness.

1. Introduction

Most large-scale assessments for grades 3-8 in the United States support both criterion- and norm-referenced score interpretations. Criterion-referenced score interpretations describe performance with respect to an external criterion domain – such as proficiency or college and career readiness – and make no direct reference to the performance of other examinees. Norm-

referencing typically addresses the question “How well did an examinee perform relative to others?” Norm-referencing can be used to describe a student’s current performance level (*status norms*) or a student’s current performance level, conditional on prior year performance (*conditional status norms*). In this study, we evaluate an imputation approach for estimating nationally-referenced status and conditional status norms.

1.1 Status norms

We refer to *status norms* as measures that are derived from marginal frequency distributions of test scores that are used to support norm-referenced interpretations. According to the *Standards for Educational and Psychological Testing*, “The validity of norm-referenced interpretations depends in part on the appropriateness of the reference group to which test scores are compared” (AERA, APA, NCME, 2014, p.97). For large-scale assessments of academic achievement, reference groups are usually defined by grade level and jurisdiction (e.g., district, state, state consortium, and national). Normative information is often reported with percentile ranks but can also include the mean, median, or other measures derived from the frequency distribution of test scores.

We use *nationally-referenced* to refer to norms that use reference groups that strive to be *nationally-representative* (e.g., as if drawn from a simple random sample of students across the nation). Nationally-referenced norms are attractive because they allow users to understand their performance relative to all other students in the United States, providing meaningful comparative data. Because different assessment systems for grades 3-8 are used throughout the United States, it is difficult to support claims of nationally-representative status norms. Special studies using random sampling techniques can be used to estimate national norms, but such studies can be difficult to implement in a way that preserves random sampling. If students or schools select into

or out of the sample (e.g., choose to participate or not participate), the resulting sample is no longer a random sample, weakening claims of national representation. Moreover, participants in special studies may not be as motivated to perform their best if the test scores don't have the same stakes as other state tests.

The National Assessment of Educational Progress program (NAEP) is a longstanding assessment of the nation's academic achievement. Among other things, NAEP produces nationally-representative estimates of mean test scores, percentiles, and percentage of students at or above certain achievement levels (NCES, 2009). NAEP uses sampling and weighting procedures to estimate nationally-representative status norms, but the validity of the estimates can be compromised by factors such as absenteeism or insufficient school participation. NAEP results have been linked to other assessment data for various purposes, such as fostering international comparisons of percentage of students performing at or above NAEP achievement levels (Phillips, 2007) and comparing proficiency standards across states (Braun & Qian, 2008). In this study, existing estimates of mean district scores based on linkages of state assessment data to NAEP are used to measure district performance using a common metric across states. As described later, the estimates of mean district scores are generally known for the population and are used for imputation.

1.2 Growth (conditional status) norms

Conditional status models (CSMs) are a family of growth models that are used to describe a student's current year performance, relative to students with similar prior year test scores (Castellano & Ho, 2013). We refer to *conditional status norms* as measures that are derived from frequency distributions of current year test scores, conditional on prior year(s) test scores. Conditional status norms support norm-referenced interpretations of student growth.

While CSMs are only one type of growth model, *conditional status* and *growth* are used interchangeably in this paper.

CSMs do not require vertical scales to support normative interpretations of student growth. Different CSMs vary by the method used to estimate the conditional score distribution. The Student Growth Percentile (SGP) model (Betebenner, 2009) is a widely-used type of CSM used to describe student growth in academic achievement from one academic year to another. The SGP model uses quantile regression to estimate the percentile rank of a student's current year score, among academic peers (students with the same prior year scores). The SGP model can be contrasted to least-squares regression, which can also be used to regress current year score on prior year scores but has restrictive assumptions such as linearity, normality, and homoscedasticity of residuals.

CSM-based normative growth measures are often reported as percentile ranks (e.g., SGPs and residual ranks from least-squares regression). Aggregate forms of CSM measures describe group-level growth and include the mean or median SGP, mean or median residual rank, and mean residual score (Castellano & Ho, 2015). Similar to status norms, CSM-based measures are interpreted with respect to the sample used to estimate the model. Because different assessment systems for grades 3-8 are used throughout the United States, CSMs usually reference students from one state or one assessment consortium. However, similar to status norms, some stakeholders may wish to interpret CSM-based measures with respect to all students nationally to enhance interpretation and comparability.

Like all growth models, CSMs require test scores from two or more points in time (typically, two or more academic years). The challenge of estimating nationally-representative CSMs is similar to the challenge of estimating nationally-representative status norms, except that

multiple years of test scores are involved and conditional score distributions (rather than marginal distributions) are of interest.

1.3 Study motivation and objectives

The study is motivated by the desire to estimate nationally-referenced status and growth norms for ACT Aspire® Summative, an assessment system that spans grades 3-10, with students generally tested at most once per year (ACT, 2017). The assessment system began in spring 2013 and the data used for this study go through spring 2017; thus, students have at most five years of test scores. From spring 2013 through spring 2017, the assessment system was adopted for grades 3-8 by one state for four years, another state for two years, and another state for one year. In addition to the states, various districts across the United States have used the assessment system for one or more years. As described later, other data used to estimate the nationally-referenced norms is only available for public schools and for grades 3-8. Thus, while the assessment system has been used by both public and non-public schools and includes grades 9 and 10, we restrict the analysis to public schools and grades 3-8.

The primary objective of the study is to evaluate how well an imputation-based method performs for estimating nationally-representative status and growth norms. As described later, a holdout method is used to examine how well the method estimates norms for a) jurisdictions that participated in the assessment program and b) for cases with intermittent missing data.

1.4 Overview of estimation method

The approach used to estimate nationally-referenced status and growth norms entails four steps:

1. *Construction of sample data set*. A longitudinal data set spanning grades 3-8 is constructed that includes public school student test scores in English, math, reading,

science, and English Language Arts (ELA). The data set also includes student gender, race/ethnicity, school characteristics, and grade-specific NAEP-referenced measures of district performance.

2. *Construction of population data set*. A longitudinal data set of the population of public school students in the United States, spanning grades 3-8, is constructed. This data set includes student gender, race/ethnicity, school characteristics, and grade-specific NAEP-referenced measures of district performance.
3. *Imputation of population data set*. After combining the sample and population data sets, imputation is used to obtain a population data set that includes imputed test scores for grades 3-8.
4. *Calculation of nationally-referenced norms*. Using the imputed population data set, summary measures of the marginal and conditional test score distributions are calculated.

Each of these four steps is now described in greater detail.

2. Methods

2.1 Construction of sample data set.

The sample data set was constructed by merging grade-specific test scores (in English, math, reading, science, and ELA) with school characteristics, student-level demographic variables (race/ethnicity and gender), and grade-specific measures of district mean achievement. Because the assessment system has been operating for five years, the sample data set contains up to five years of test scores for grades 3-8, with test scores missing intermittently. Students were included in the sample data set if they tested at least once in math and reading, if they were enrolled in a public school with known NCES school identifier code, and if they were in a district that met the district inclusion criteria discussed later. Most students (51%) had one year of assessment data, 26% had two years, 10% had three years, 12% had four years, and 1% had

five years. By grade level, the number of students tested ranged from 310,550 for grade 3 to 359,761 for grade 7. The overall sample size was 1,091,534.

The school characteristics data was obtained from the National Center for Education Statistics (NCES) Common Core of Data for 2014-2015 (Glander, 2016) and included enrollment count, percentage of students eligible for free or reduced lunch (FRL), percentage of African American and Hispanic students, locale (rural, town, suburban, or urban), and geographic region (Midwest, Northeast, South, or West).

2.1.1 Measures of district mean achievement

The Stanford Education Data Archive (SEDA; Reardon, Ho, Shear, Fahle, Kalogrides, & DiSalvo, 2017) includes measures of academic achievement and achievement gaps for virtually all public school districts in the United States. SEDA is intended to provide researchers with data "to generate evidence about what policies and contexts are most effective at increasing educational opportunity, and that such evidence will inform educational policy and practices" (Fahle, Shear, Kalogrides, Reardon, DiSalvo, & Ho, 2017). We use SEDA's district achievement data to help estimate nationally-referenced status and growth norms.

Version 2.0 of SEDA contains estimates of district achievement (means and standard deviations) for grades 3-8 in ELA and math for 2009 (2008-2009 school year) through 2015 (2014-2015 school year). Estimates are provided for all students, as well as racial/ethnic subgroups (Asian, Black, Hispanic, and White). The SEDA achievement data is constructed using data from each state's standardized testing program for grades 3-8 as required by federal law. Coarsened data for each district (e.g., percentage of students in each achievement level) is transformed to means and standard deviations using various forms of ordered probit models (Reardon, Shear, Castellano, & Ho, 2017). The state-referenced estimates are then placed on the

NAEP scale so that they are comparable across states, years, and grades (Reardon, Kalogrides, & Ho, 2017). Each state's math scores are linked to the NAEP math scale, and each state's ELA scores are linked to the NAEP reading scale. We refer the interested reader to SEDA's technical documentation (Fahle et. al, 2017) for full details on the methodology used to generate the district statistics and for more details on other data available from SEDA.

Because the district achievement data are central to our method for estimating nationally-referenced norms, we only include students who belong to a district included in the SEDA data. SEDA provides a crosswalk file of NCES school codes to SEDA's geographic school districts (GSDs), and it was used to map each student's records to a district, consistent with how data are grouped in SEDA. For each grade level and year combination, we required that a district's sample size (in our sample data set) was within 10% of the district's sample size used for the SEDA estimates. This ensures that both the sample data set and SEDA district achievement statistics account for virtually all students in the district.

For each record in the sample data set, the district mean and standard deviation of math and ELA scores is obtained from SEDA, along with the mean math and ELA scores for the racial/ethnic subgroups. The sample data set includes records from 2013 through 2017, while the SEDA data covers 2009 through 2015. For sample data from 2013 through 2015, we used the SEDA data from the same year. For 2016 and 2017, estimates of the SEDA data were obtained using linear extrapolation of the 2009-2015 data. The extrapolated estimates of the all-student statistics were then used for the sample data for 2016 and 2017. In addition to the year-specific SEDA data, pooled SEDA data was estimated by averaging across years within a district and grade level. The pooled SEDA estimates were used in cases where the year-specific data was not

otherwise available. The racial/ethnic subgroup statistics for 2016 and 2017 were based on the pooled SEDA estimates.

2.2 Construction of population data set

The population data set is designed to represent the population of public school students in the United States, while that population progresses through grades 3-8. The actual population varies somewhat by year, and we chose 8th graders of 2015 (2014-2015 school year) as the basis for creating the population data set. For each public school, the NCES data contains the number of 8th-grade students by gender and race/ethnicity. The NCES data also include other school characteristics that can potentially be useful for imputation (the same variables listed earlier when describing the sample data). Similar to the sample data set, the NCES data can be linked to the SEDA data set using NCES school code. While we began with the 8th-grade population as the basis for the population, we are interested in imputing the population's test scores for grades 3-8, and so we obtained the SEDA data for grades 3-8. The SEDA data for each grade level was obtained by linking to the appropriate year of SEDA data, assuming that students in the 8th-grade population had progressed one grade level each year (2015 for grade 8, 2014 for grade 7, etc.). Following these steps, the population is defined with respect to student ethnicity and gender, school characteristics, and district mean achievement for grades 3-8.

The population data set has the same variables as the sample data set, except for the student test scores. Table 1 compares the sample to the population on student demographics and school characteristics. The sample consists of about 1.1 million students who tested between 2013 and 2017 while in grades 3-8, so it contains students in several 8th-grade cohorts (2013 through 2022). The population consists of over 3.7 million students and uses the 2015 cohort of 8th-grade students as its basis. The sample is similar to the population on gender, but contains

more African American students and fewer Hispanic and Asian students. The sample is mostly based in the South region of the United States (~94%), with very little representation from the Northeast and West regions. Schools located in rural and town settings are over-represented in the sample, relative to schools in suburban and urban settings. The percentage of students eligible for free or reduced lunch is similar for the sample (56%) and population (52%).

Table 1. Comparing Sample to Population

Variable	Sample	Population
N students	1,091,534	3,725,600
N districts	597	12,910
Gender		
Female	49.4	48.7
Male	50.4	51.2
Missing	0.3	<0.1
Race/ethnicity		
African American	27.6	15.6
Asian	1.5	4.9
Hispanic	8.2	24.7
Other	2.6	4.1
White	54.0	50.6
Missing	6.1	<0.1
Census region		
Midwest	3.9	21.0
Northeast	0.0	15.8
South	94.1	39.0
West	2.0	24.2
School locale		
Rural	35.0	18.9
Town	15.1	11.3
Suburban	25.8	40.1
Urban	23.1	29.2
Missing	1.1	0.6
School FRL% (mean)	55.8	51.6

2.3 Imputation of population data set.

After concatenating (stacking) the sample and population data sets, imputation is used to obtain a population data set that includes imputed test scores for grades 3-8. The population data

set is missing all test scores, while the sample data set has intermittently missing test scores. The SAS MI procedure (SAS, 2011) is used to impute a complete data set. The MI procedure assumes that data are *missing at random*, meaning that missingness may depend on the observed data, but not the missing data (Rubin, 1976). The MI procedure also assumes that the parameters of the data model are distinct from the parameters of the missing data model. This means that knowing the parameters of the data model does provide information about the parameters of the missing data model, and vice-versa (Yuan, 2011).

The appropriate imputation method used with the MI procedure generally depends on the pattern of missingness (monotone or arbitrary) and the type of imputed variable (continuous, ordinal, or nominal; Yuan, 2011). A monotone missing pattern occurs when missing a variable implies that all subsequent variables are also missing. Our data set has arbitrary missingness, contains mostly continuous variables (student test scores, all-student district mean achievement, racial/ethnic group-specific district mean achievement, school FRL%, school minority %, school enrollment), and contains three nominal variables (race/ethnicity, gender, and school locale). For data sets with arbitrary missingness and continuous variables, the recommended MI method uses Markov Chain Monte Carlo (MCMC) methods to simulate the joint posterior distribution of the observed and missing data. Each variable specified in the imputation model informs the imputation of all other variables.

Using the SAS MI procedure with the MCMC method assumes that the data are drawn from a multivariate normal (MVN) distribution, so it is not appropriate for nominal variables. While our data set contains three nominal variables, we chose to use the MCMC method anyway, reasoning that the nominal variables (coded as dummy variables) are essentially used as covariates and the imputed values of the nominal variables are not important. SAS MI also

supports predictive mean matching (PMM; Little, 1988), which does not assume MVN but instead draws imputed values from observed values and preserves the original shape of the frequency distribution. However, the missing data must be monotone to use PMM. The MICE (Multiple Imputation by Chained Equations) package (van Buuren & Groothuis-Oudshoorn, 2011) also supports PMM, among other methods, and does not require monotone missing data. However, using the MICE function with PMM on our large data set was not feasible because it took too long to run. Later, we discuss how the MVN assumption affects the imputation of the student test score distributions, which do not generally conform to MVN.

Because imputation adds an additional random component to the data, analyses of multiple imputed data sets is generally recommended (Yuan, 2011). We imputed five data sets.

2.4. Calculation of nationally-referenced norms.

Using the imputed population data set, nationally-referenced norms are calculated. Because the imputed data set includes complete vectors of test scores for grades 3-8, both grade-specific marginal frequency distributions (for status norms) and joint frequency distributions (for conditional status norms) are available. For each grade level (3-8) and two subject areas (math and reading), the frequency distribution of imputed test scores is obtained. Means, standard deviations, selected percentiles (e.g., 5th, 25th, 50th, 75th, and 95th), and cumulative percentile ranks (corresponding to each score) are computed from the frequency distributions.

Using the joint frequency distributions, CSMs such as the SGP model and multiple linear regression (MLR) can be fit to the data to produce CSM-based measures such as SGPs and ranks of residual scores. We found that the sample's residual ranks estimated using MLR corresponded closely with SGPs derived using the SGP R package[1] (Betebenner, VanIwaarden, Domingue, &

[1] We used the studentGrowthPercentiles function with default settings.

Shang, 2017), with a median correlation of 0.993 across the ten subject area/grade level combinations (min=0.987, max=0.996). While the SGP model is preferred over MLR because of MLR's restrictive assumptions, we focus on MLR-based residual ranks in this study because the models produce similar results and because the MLR model is easier to fit using the software used for imputation (SAS). Current-year test scores are regressed on the prior year score (linear and quadratic effects) in the same subject area.

2.5 Evaluation of estimation method

The imputation-based estimation method produces nationally-representative status and growth norms if the imputed marginal and conditional score distributions match the actual distributions. The method may work well if data are missing at random (e.g., if the probability of missing test scores depends on student demographics, school characteristics, and/or district mean achievement statistics, but not on other unobserved variables) and if the unobserved test scores conform to MVN. Neither of these conditions is likely to be strictly met.

While we cannot prove that the population norms are nationally representative, we can examine how well the method recovers marginal and conditional distributions for students and jurisdictions that participated in the assessment program. If the method works well for members of the population that participated in the assessment program, it may be reasonable to expect it to work well for members of the population that did not. Using a holdout approach, imputation accuracy is examined for students and jurisdictions in the sample data set. Our study addresses two types of missing data – system missingness (e.g., all students in a jurisdiction are missing test scores because the jurisdiction did not participate in the assessment program) and intermittently missing test scores (e.g., a student's test scores are observed for grades 3 and 4, but not 5-8).

2.5.1 Evaluation of estimation method for system missingness

The sample data set was divided into seven jurisdictions. Recall that the assessment program has been used for grades 3-8 by three states and by individual districts in other states. We divided the three states into six jurisdictions according to locale (urban/suburban or rural/town). These six jurisdictions can this be thought of as "mini states," and defining jurisdictions in this manner allows us to evaluate how the estimate procedure would recover score distributions for states that have not participated in the assessment program. The seventh jurisdiction represents the other districts from various states.

The overall jurisdiction sample sizes ranged from 75,985 to 252,723. For each jurisdiction, the following steps are taken to examine accuracy of imputation for system missingness:

1. For all students in the jurisdiction, set the test score data to missing in the sample data set (holdout sample). Concatenate the sample and population data sets to form the total data set.
2. Impute data for the total data set. This step imputes test score data for the jurisdiction of interest, the population, and the intermittent missing test scores in the sample.
3. Generate two sets of status norms for the jurisdiction of interest. The first set of norms is based on actual test scores that were held out in step 1. Let $f_h(Y)$ represent the marginal score distribution for the holdout sample, with cumulative frequency distribution (CFD) $F_h(Y)$. The second set of norms is based on the imputed test score data for the jurisdiction. Let $g_h(Y)$ represent the marginal score distribution for the holdout sample, with CFD $G_h(Y)$.

4. Compare the jurisdiction's imputed and actual status norms. The following measures are calculated:
    a. d = (imputed mean – actual mean) / actual SD.
    b. d_p (for p = 5, 25, 50, 75, 95) = (imputed p^{th} percentile – actual p^{th} percentile) / actual SD.
    c. SD ratio = imputed SD / actual SD.
    d. Frequency distribution of D_{rank}, where $D_{rank,\, i} = G_h(i) - F_h(i)$ represents the difference in the CFD at the ith test score. D_{rank} is assigned to each observed test score in the jurisdiction's data set and represents the difference in the percentile ranks that would be assigned using the imputed data set and using the actual data set.
5. Generate two sets of conditional status norms for the jurisdiction of interest. The first set of norms is based on the actual test scores that were held out in step 1. For each grade level pair (3-4, 4-5, etc.) and each subject, MLR is used to produce residual ranks. Similarly, MLR is applied to the imputed data set to produce a residual rank for each combination of test scores. The residual ranks from the imputed data set are then applied back to the actual data set so that it contains two residual ranks (one based on the MLR model fit using actual data set, one based on the MLR model fit with the imputed data set).
6. Compare the jurisdiction's imputed and actual conditional status norms. We are primarily interested in the distribution of D_{gp}, the difference in growth percentiles (residual ranks) based on the imputed and actual data sets. We calculate the mean, standard deviation, 5th percentile, and 95th percentile of $D_{gp.}$

Steps 1 through 6 for evaluating imputation accuracy are repeated for different jurisdictions used as the holdout sample (7), subject areas (2), and grade levels (6), providing up to 84 (7*2*6) comparisons of status norms based on imputed data or actual data. To minimize differences between imputed and actual status norms that is due to sampling variability, we require that the actual data set have at least 5,000 observed test scores for the jurisdiction of interest (for each combination of grade level and subject area). We are left with 80 of the 84 comparisons. For examining conditional status norms, there are up to 70 comparisons (7 jurisdictions, 2 subject areas, and 5 grade level pairs). We required at least 5,000 observed test score pairs, leaving us with 68 comparisons for examining the accuracy of imputed conditional status norms.

2.5.2 Evaluation of estimation method for intermittent missingness

The following steps are taken to examine accuracy of imputation for intermittently missing test scores:

1. Consider students with observed test scores in math and reading for at least two grade levels. Randomly select 50% of the students to have their test scores for one grade held out. For each selected student, randomly select the grade to hold out and set all test scores for that grade to missing. Concatenate the sample and population data sets to form the total data set.
2. Impute data for the total data set. This step imputes test scores for the population and all intermittent missing test scores in the sample (including those that were set to missing in step 1).
3. For the holdout data, compare the imputed and actual test score distributions. The following measures of imputation accuracy are calculated:

a. d = (imputed mean – actual mean) / actual SD.

b. SD ratio = imputed SD / actual SD.

c. Mean of D_{rank} and mean of $|D_{rank}|$ where D_{rank} is the difference in the percentile ranks that would be assigned using the imputed data set and using the actual data set.

3. Results

3.1 Estimates of nationally-referenced status norms

First, we describe features of the nationally-referenced status norms that are generated using the process described in sections 2.1-2.4. The district mean NAEP-referenced scores are available for both the population and sample. Figure 1 shows frequency distributions of grade 8 district mean math NAEP-referenced scores for the sample and population. While the population's scores are higher than the sample's, there is considerable overlap in the distributions.

Figure 2 provides mean NAEP-referenced math scores for the population, sample before imputation, and sample after imputation (after imputation, the sample has no intermittent missing data). Across all grade levels, the mean NAEP-referenced math scores are lower for the sample relative to the population. Similarly, Figure 3 provides mean NAEP-referenced ELA scores for the population, sample before imputation, and sample after imputation. The mean NAEP-referenced ELA scores are similar for the sample and population.

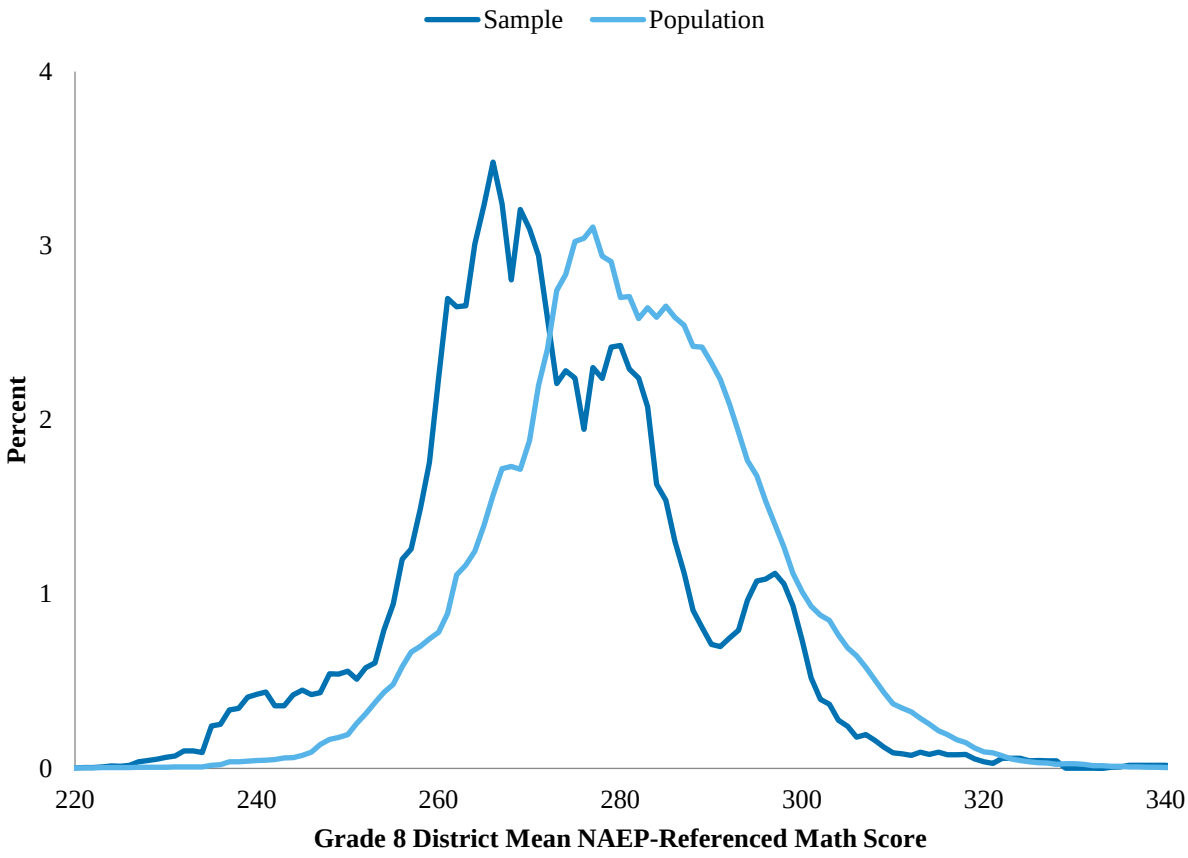


Figure 1. Distributions of grade 8 district mean NAEP-referenced math scores

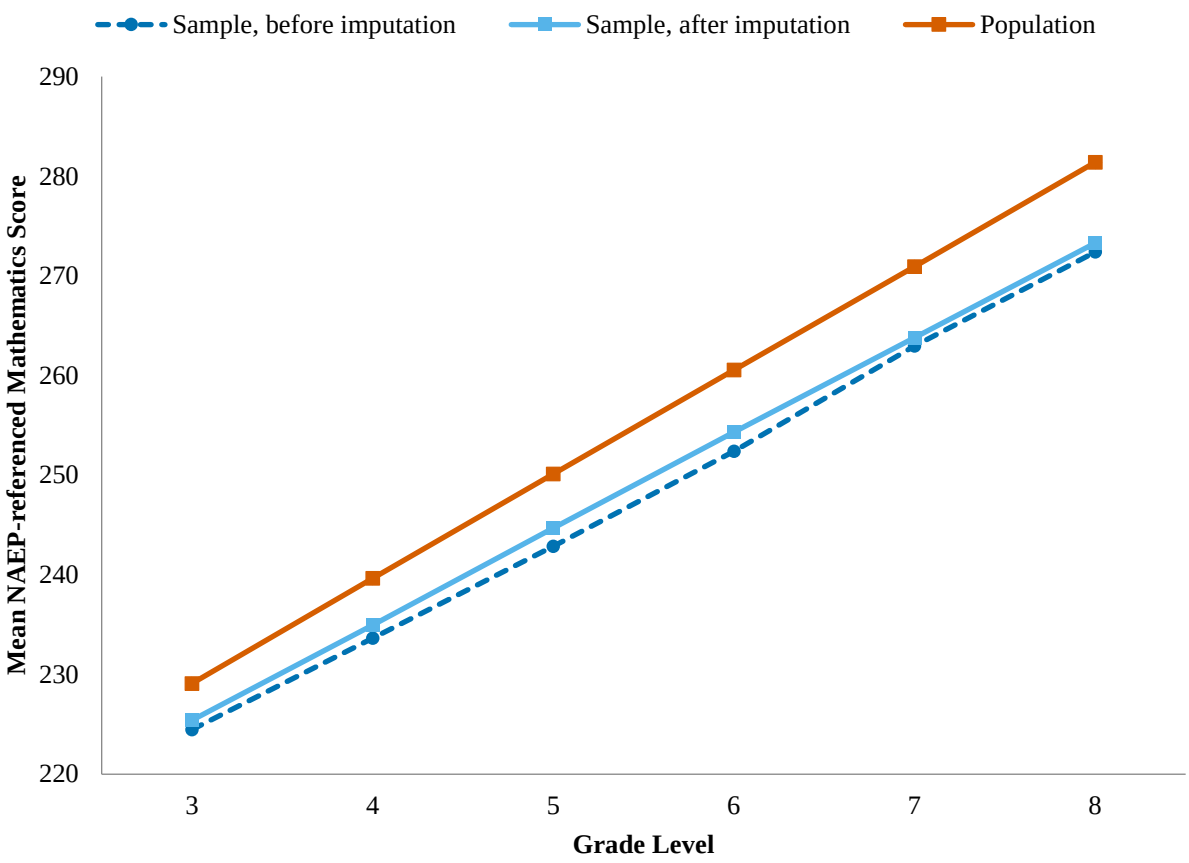


Figure 2. Mean NAEP-referenced math scores

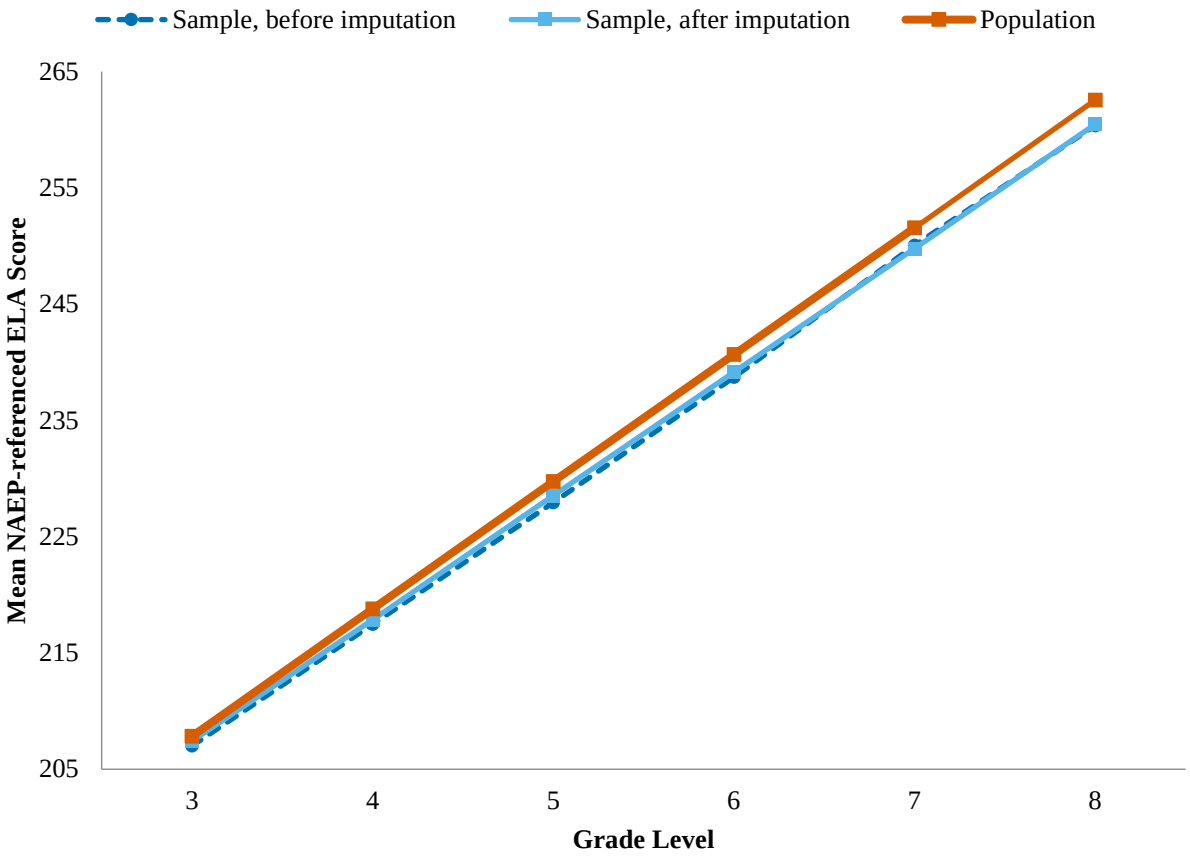


Figure 3. Mean NAEP-referenced ELA scores

Because district mean math scores are higher in the population than in the sample, we would expect the student test scores to be higher in the population than in the sample. Figure 4 shows that this is the case: Mean ACT Aspire math scores for the population (based on imputation) are higher than mean ACT Aspire math scores for the sample. Similarly, we would expect the mean student test scores in reading to be similar in the population and sample, and Figure 5 shows that this is the case.

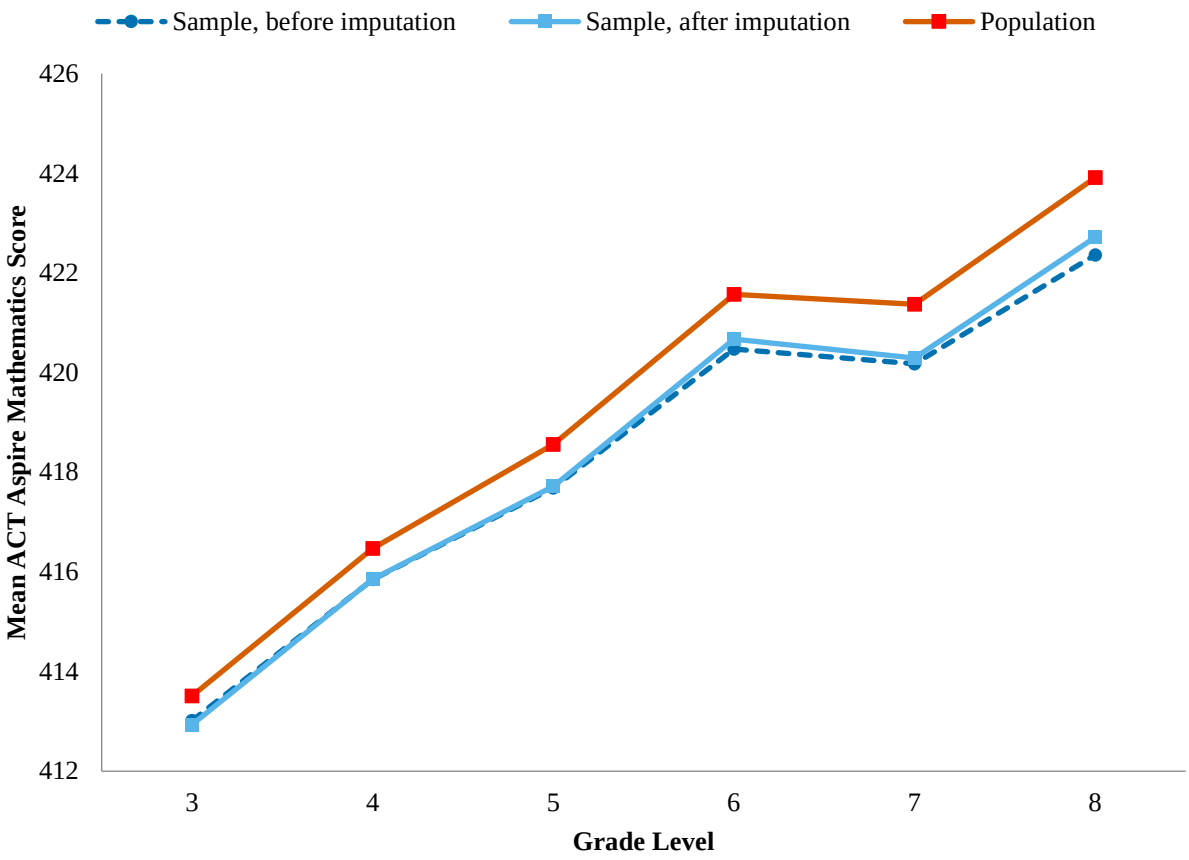


Figure 4. Mean ACT Aspire math scores

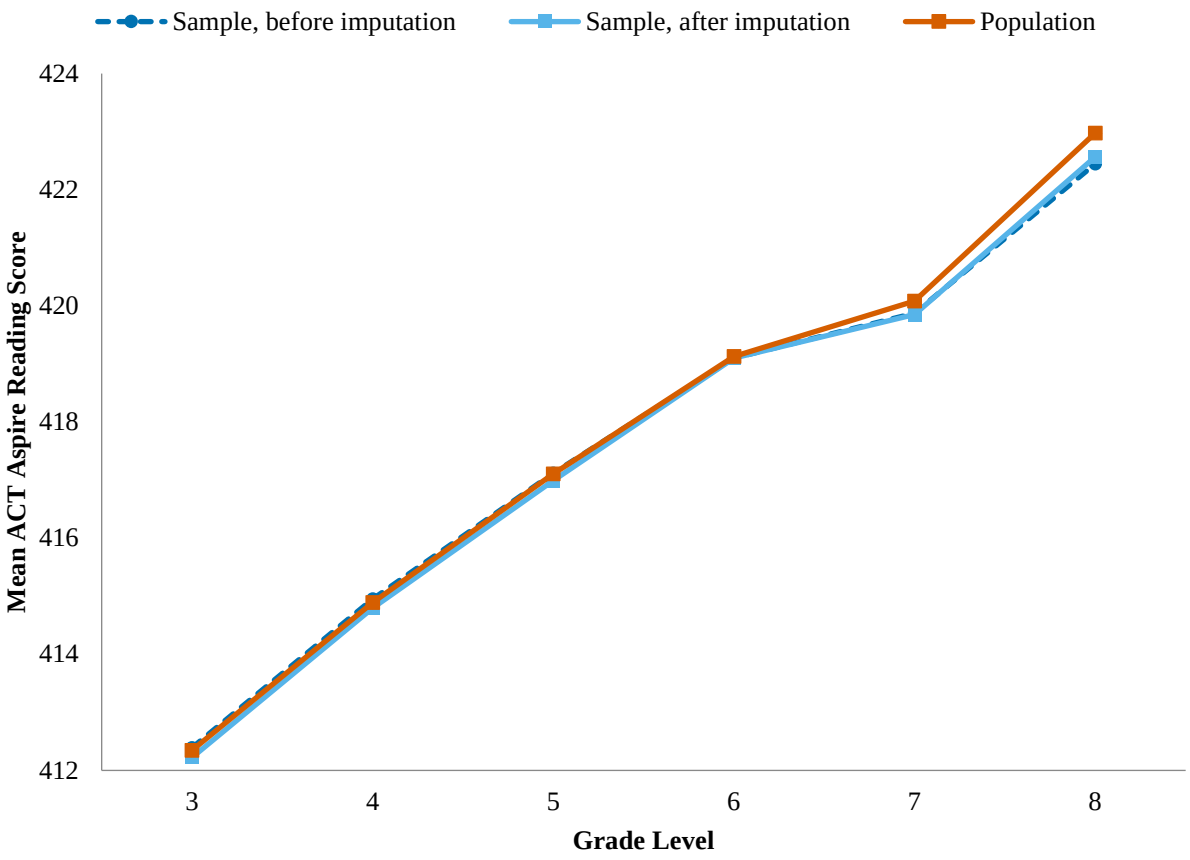


Figure 5. Mean ACT Aspire reading scores

In Figures 2 and 3, the mean NAEP-referenced scores progress in a linear fashion across grade levels. NAEP is administered only in grades 4 and 8 in odd years (e.g., 2009, 2011, 2013, 2015). SEDA's NAEP-referenced estimates of district mean achievement are linked to each state's mean NAEP scores, which are based on linear interpolation (for grades 5-7 and years 2010, 2012, and 2014) and extrapolation (for grade 3; Fahle et al., 2017). The linear relationships observed in Figures 2 and 3 are an artifact of the linear interpolation/extrapolation. The population mean NAEP-referenced scores for grade 8 (281.4 for math, 262.6 for ELA) should be similar to the 2015 national public school estimates from NAEP for grade 8 (281 for math, 264 for reading; The Nation's Report Card, 2017). Similarly, the population mean NAEP-referenced

scores for grade 4 (239.7 for math, 218.8 for ELA) should be similar to the 2011 national public school estimates from NAEP for grade 4 (240 for math, 220 for reading; The Nation's Report Card, 2017). In Figures 4 and 5, scores are increasing across grade levels, with the exception of grade 6-7 math. Because ACT Aspire is vertically-scaled, scores are expected to increase, though not necessarily linearly.

Figures 2-5 also show that imputing the intermittently missing sample data tends to result in a small increase in mean test scores. The NAEP-referenced means show that the gap between the sample and population is increasing across grade levels, suggesting that growth in the sample is lower than that observed nationally. For example, in Figure 2, the sample mean math score is 0.12 standard deviations below the population mean for grade 3 but 0.22 standard deviations below the population mean for grade 8.[2] In Figure 4, the sample mean math score is 0.13 standard deviations below the population mean for grade 3 and 0.15 standard deviations below the population mean for grade 8. The standard deviation of 2015 NAEP math scores increases somewhat from grade 4 (30) to grade 8 (37), while the standard deviation ACT Aspire math scores increases more drastically from grade 4 (4.3) to grade 8 (8.2). In Figure 4, the gap between the sample and population appears to increase across grade levels. However, this is mostly due to the increasing standard deviation: In standard deviation units, the gap is very consistent across grade levels.

The frequency distributions of grade 3 ACT Aspire math test scores are provided in Figure 6. The distribution for the population is shifted to the right of the sample distributions, resulting in the larger mean shown in Figure 4. Because the imputation software assumes that test scores are distributed as MVN, the imputed population frequency distribution is normally

[2] Estimates of the standard deviations of 2015 NAEP math scores are 30 for grade 4 and 37 for grade 8.

distributed. While the sample distribution is not normal, the imputed sample distribution looks more like a normal distribution because intermittent missing values are imputed as MVN.

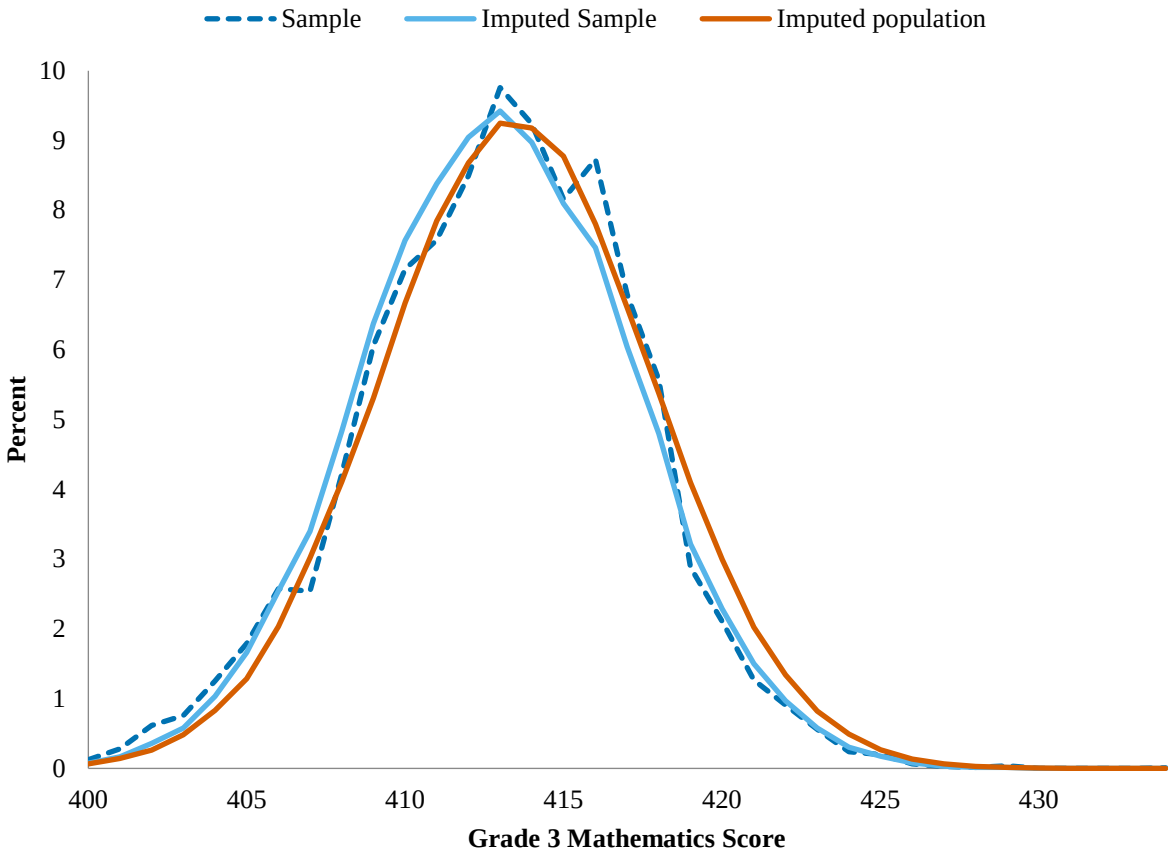


Figure 6. Frequency distributions of grade 3 ACT Aspire math scores

For math, the estimated population mean was consistently larger than the sample mean (after imputation), ranging from *d*=0.13 for grade 3 to *d*=0.16 for grade 5. For reading, the estimated population means were only slightly larger than the sample mean, ranging from 0.005 for grade 6 to 0.06 for grade 8. After imputation, the population standard deviations were slightly larger than the sample standard deviations. Across the two subject areas and six grade levels, the ratio of sample to population standard deviation ranged from 0.975 to 0.993.

3.2 Estimates of nationally-referenced conditional status norms

For each grade level pair and both subject areas, MLR was applied to the sample data (after imputation) and imputed population data set to obtain residual ranks. The residual ranks estimated using the population data were then applied back to the sample data set so that we could assess differences in sample and population-based residual ranks.

The largest differences in sample and population-based conditional status norms were observed for grade 4-5 math, where the sample-based residual ranks were 2.1 points higher than the population-based residual ranks, on average. In this case, the sample would overestimate a student's growth percentile, relative to the growth percentile they would be assigned if based on the population data. However, the direction of the difference in residual ranks varied by grade 4 math score. Figure 7 shows the predicted grade 5 math scores, by grade 4 math score.[3] The solid lines represent the predicted values, and the dashed lines represent the upper and lower bounds of 75% prediction intervals. For very low and very high grade 4 math scores, the sample-based predictions are higher than the population-based predictions. In these cases, the sample-based residual ranks are lower than the population-based residual ranks. For grade 4 scores in the middle of the distribution, the sample-based residual ranks are higher than the population-based residual ranks.

In Figure 7, we see that the sample's regression lines are curvilinear, indicating that the quadratic effect was important, while the population's lines are more linear. The population estimates are based on imputed data that was drawn from an MVN distribution, while the sample estimates are based on actual test scores. Further examination is needed to determine if the assumption of MVN forces linearity in the population's regression lines.

[3] Grade 4 scores from the 1st to 99th percentile are shown.

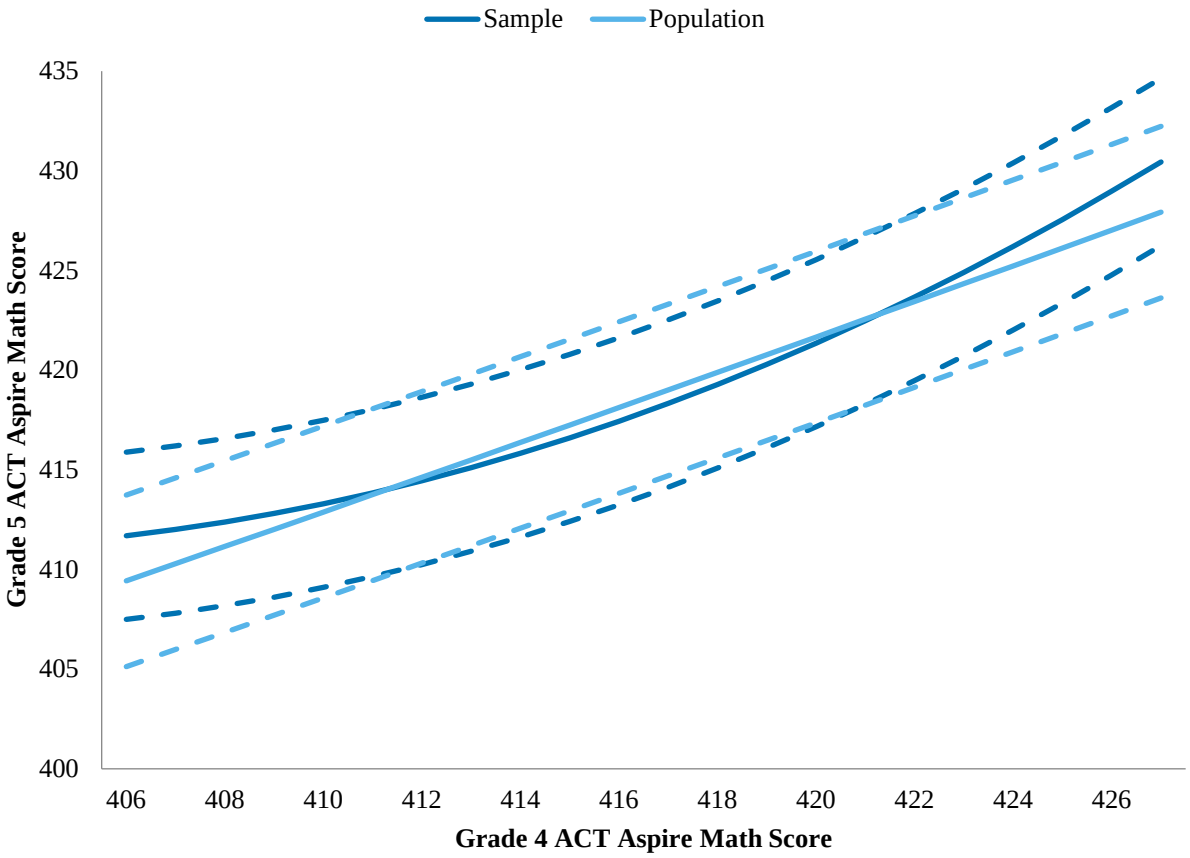


Figure 7. Sample and population-based regression results for grade 5 math

Note: Solid lines are predicted values and dashed lines represent lower and upper bounds of 75% prediction interval.

Figure 7 (grade 4 to 5 math) demonstrates the largest difference (across grade levels and subjects) between sample and population conditional status norms. Figure 8 (grade 4 to 5 reading) shows the smallest difference. Estimates of conditional status norms for grade 4 to 5 reading are very similar for the sample and population, with the predicted values and prediction intervals hard to distinguish.

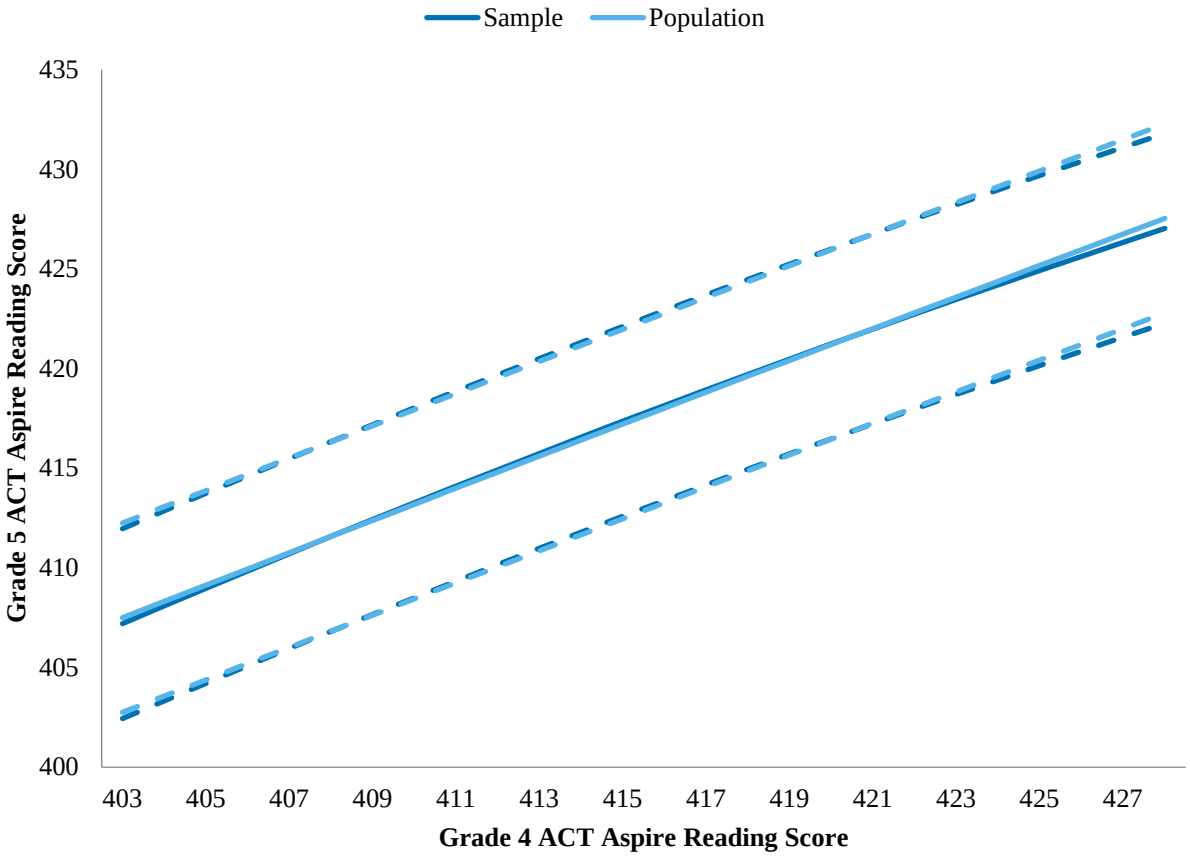


Figure 8. Sample and population-based regression results for grade 5 ACT Aspire reading

Note: Solid lines are predicted values and dashed lines represent lower and upper bounds of 75% prediction interval.

Across all grade levels and subjects, the sample estimates of residual ranks tended to be slightly higher than the population-based estimates, with a mean difference of 0.55 percentile points (0.54 for math, 0.56 for reading). For example, if a school's sample-referenced mean growth percentile was 49.0, we would expect its population-referenced mean growth percentile to be about 48.5. This suggests that students in the sample grow slightly less than students in the population.

3.3 Accuracy of status norms under system missingness

The accuracy of the status norms was assessed across the seven jurisdictions that were successively held out of the sample data set. As described earlier, there are 80 comparisons of imputed and actual marginal frequency distributions (7 holdout jurisdictions*6 grade levels*2 subject areas, less 4 comparisons with N < 5,000). On average, the imputed mean was 0.01 standard deviations below the actual mean (*d*=-0.01). Thus, on average, the imputation procedure came very close to correctly estimating the mean. However, this result is expected because jurisdictions are successively held out: If the mean is over-estimated for one jurisdiction, we would expect it to be under-estimated for other jurisdictions. Thus, *d*=-0.01 is not evidence of the accuracy of the imputation procedure.

At most, the imputed mean was 0.18 standard deviations below the mean (min *d*=-0.18) and 0.21 standard deviations above the mean (max *d*=0.21). Across the 80 comparisons, the 5th percentile of *d* was -0.15 and the 95th percentile was 0.16. For 79% of the comparisons, the imputed mean was within 0.10 standard deviations of the actual mean.

There was some evidence that the estimation method did not work as well in the tails of the distribution. On average, the imputed 5th percentile was 0.10 standard deviations below the actual 5th percentile. Similarly, the imputed 95th percentile was 0.04 standard deviations below the actual 95th percentile. The percentage of comparisons that the imputed percentile was within 0.20 standard deviations of the actual percentile was 73% for the 5th percentile, 91% for the 25th percentile, 89% for the median, 84% for the 75th percentile, and 76% for the 95th percentile.

Across the 80 comparisons, the imputed standard deviation was 1% lower than the actual standard deviation on average (min=11% lower, max=11% higher). For 96% of the comparisons,

the imputed standard deviation was within 10% of the actual standard deviation; for 71% of the comparisons, the imputed standard deviation was within 5% of the actual standard deviation.

Percentile ranks were estimated from the imputed data and actual data set, and then applied back to the actual data set. D_{rank} is the difference in percentile ranks (rank based on imputed data – rank based on actual data). D_{rank} ranged from -10 to 13, with a 5th percentile of -5, 25th percentile of -2, median of 0, 75th percentile of 2, and 95th percentile of 7 (Figure 9). In 99% of cases, percentile ranks estimated from imputed data were within 10 points of those estimated from the actual data; 85% were within 5 points and 69% were within 3 points.

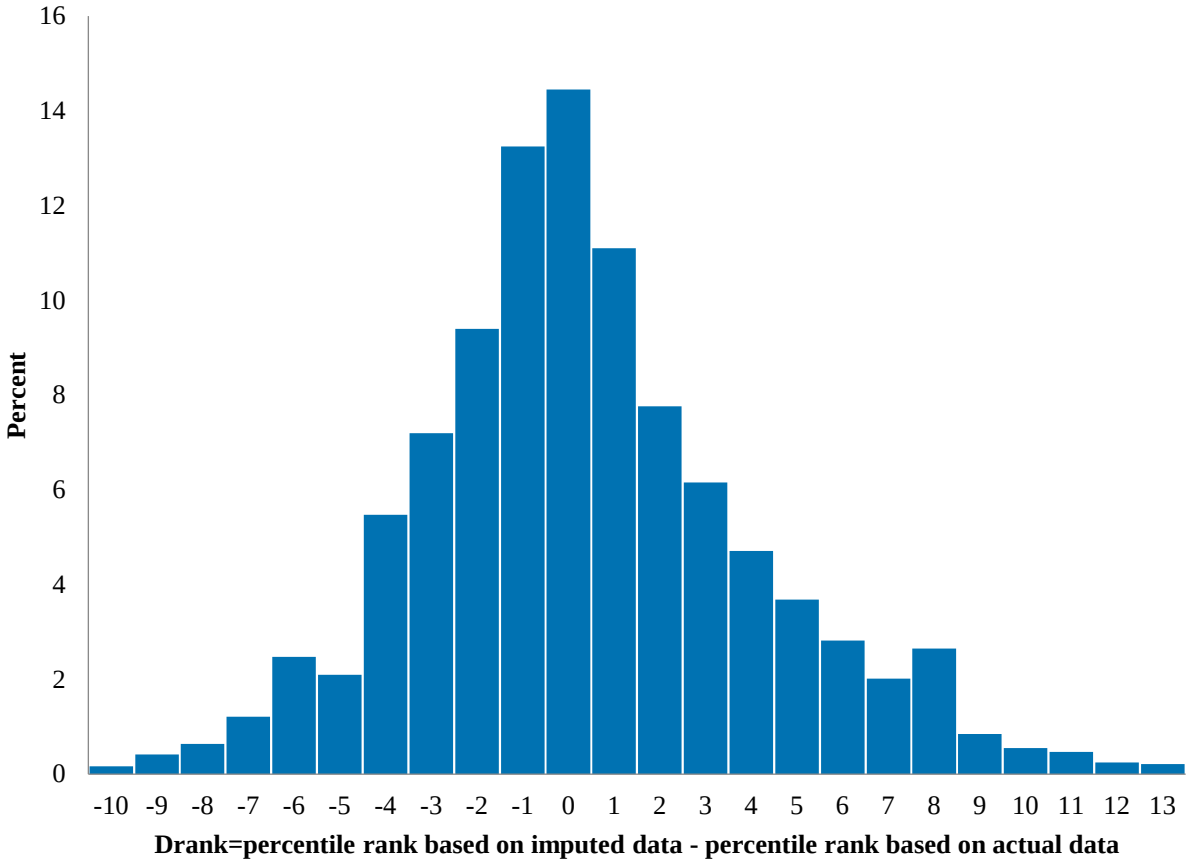


Figure 9. Distribution of differences in percentile ranks based on imputed and actual data

3.4 Accuracy of conditional status norms

Residual ranks estimated using MLR were estimated from the imputed data and actual data, and then applied back to the actual data. d_{gp} is the difference in growth percentiles (residual ranks) based on the imputed and actual data sets. d_{gp} ranged from -81 to 20, with a 5th percentile of -6, 25th percentile of -1, median of 0, 75th percentile of 2, and 95th percentile of 5. In 92% of cases, the residual rank estimated from imputed data was within five points of that estimated with actual data; 79% of cases were within three points; and 54% were within one point.

Figure 10 illustrates a comparison of conditional status norms based on imputed and actual data (one jurisdiction, grade 6-7, math). Figure 10 shows the predicted grade 7 math scores, by grade 6 math score. The solid lines represent the predicted values, and the dashed lines represent the upper and lower bounds of 75% prediction intervals. For very low grade 6 math scores, the predictions based on the actual data are higher than the predictions based on the imputed data. In these cases, residual ranks based on actual data are lower than those based on the imputed data. For very high grade 6 math scores, the opposite is true. For grade 6 scores in the middle of the distribution, there is little difference between residual ranks based on imputed and actual data. Overall, for this comparison, 98% of the students have a residual rank based on the imputed data that is within 2 points of the residual rank based on the actual data.

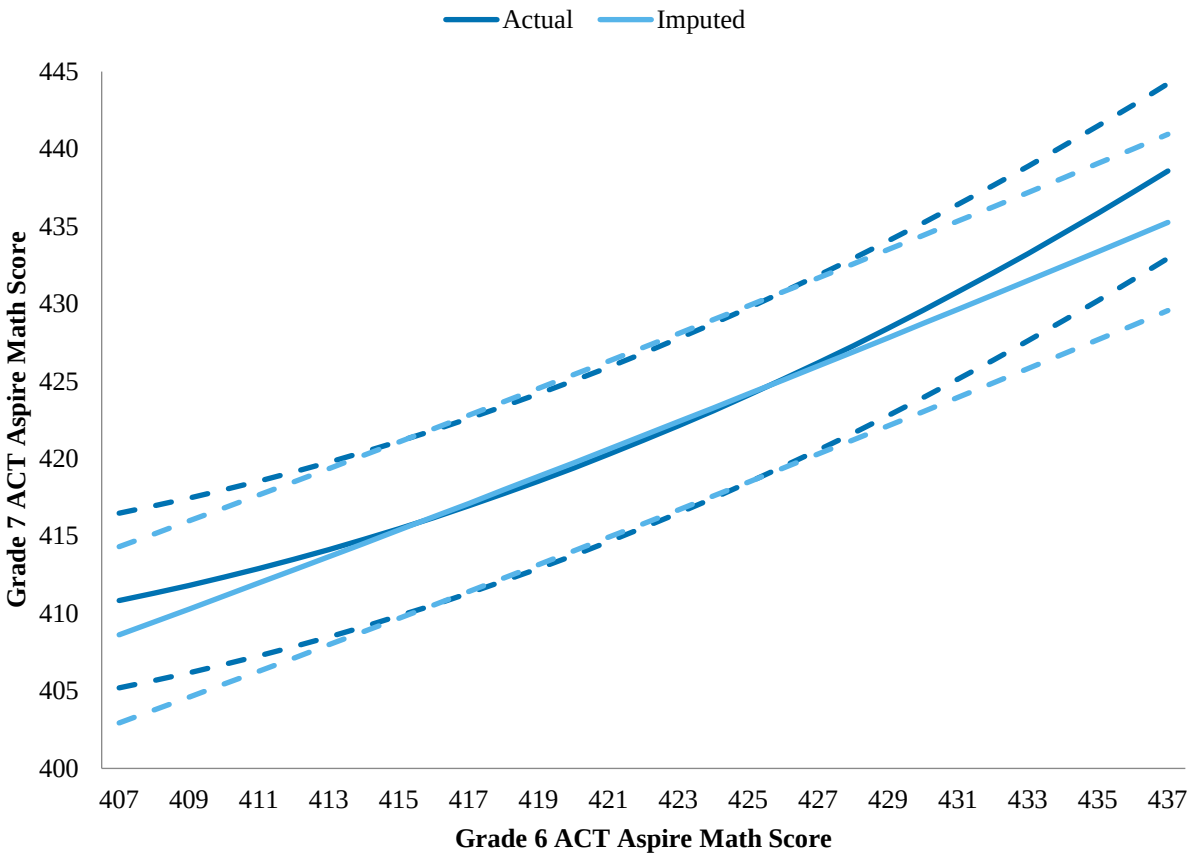


Figure 10. Regression results for grade 7 math based on actual and imputed data for one jurisdiction

3.5 Accuracy of imputation for intermittent missingness

To examine the accuracy of imputing intermittently missing test scores, scores from one grade level were held out for half of the students who tested in at least two grade levels. Imputation for intermittent missingness was very accurate. Across the six grade levels and two subject areas, the imputed mean was at most 0.03 standard deviations from the actual mean (Table 2). For grade 8 math, the imputed mean was 0.03 standard deviations below the actual mean.

Table 2. Accuracy of Imputation for Intermittently Missing Data

Subject	Grade	N	*D*	SD ratio	Mean D_{rank}	Mean $\lvert D_{rank} \rvert$
Math	3	34,950	0.02	1.02	0.09	1.30
	4	50,678	-0.01	1.01	-0.20	0.88
	5	46,305	0.02	1.00	-1.57	2.11
	6	46,175	0.00	1.01	-0.54	0.95
	7	52,314	-0.01	1.01	-0.39	1.25
	8	37,237	-0.03	0.99	-0.26	1.79
Reading	3	34,950	0.00	1.00	-0.90	2.04
	4	50,678	-0.01	0.99	-0.02	1.05
	5	46,305	0.01	0.99	-0.74	1.12
	6	46,175	0.00	0.99	0.13	1.31
	7	52,314	0.00	1.00	0.31	1.68
	8	37,237	0.01	1.00	0.28	1.55

Note: N = number of test scores held out and used to examine imputation accuracy

Similarly, the imputed standard deviations were very close to the actual standard deviations. The largest discrepancy occurred for grade 3 math, where the imputed standard deviation was 2% higher than the actual standard deviation. Percentile ranks were assigned to the actual data according to the actual frequency distribution and according to the frequency distribution of imputed data. The largest difference in mean percentile rank occurred for grade 5 math, where percentile ranks based on the imputed data set were 1.57 points lower than those based on the actual data set. On average, the absolute difference in grade 5 math percentile ranks was 2.11.

4. Discussion

Nationally-referenced status and growth norms were estimated using a large sample of students who tested at least once in math and reading in grades 3-8. Nationally-referenced norms are desirable because they communicate how well a student or school performed with respect to a reference group that is meaningful for all jurisdictions. When the states and districts that participate in an assessment program change year-by-year, nationally-referenced norms can

provide greater stability in norms over time. Nationally-referenced norms for grades 3-8 also support vertically-moderated standard setting (Huynh & Schneider, 2005). By imputing intermittently missing test scores, we ensure that the sample is constant across grade levels and performance standards can be set consistently across grade levels.

We demonstrated an imputation-based analytic approach for estimating nationally-referenced norms. In contrast to design-based (sampling) methods, our approach is less costly because it doesn't involve recruitment of schools, adherence to sampling protocols, additional testing, or special data collection. The tradeoff is that our method assumes that data are missing at random and that the missing data model is the same for assessment participants and nonparticipants. Because the missing at random assumption can't be verified statistically, we have no guarantee that the procedure performed well for jurisdictions that did not participate in the assessment program. So the procedure can be used to produce nationally-referenced norms, but there is no guarantee that the norms are nationally representative.

Relative to the sample, the population estimates were higher in math and reading, though the reading differences were very small. Growth estimates for the population were greater than those for the sample, though the growth differences were very small. Using a holdout approach, we found that the imputation-based estimation procedure performed well for jurisdictions that participated in the assessment program, especially with respect to conditional status norms. The procedure performed very well for recovering the distributions of intermittently missing data.

The results suggest that nationally-representative norms are easier to achieve for conditional status norms, relative to status norms. One possible reason for this is that there is more variation across jurisdictions (e.g., districts) in status (e.g., mean percentile ranks) than conditional status (e.g., mean residual ranks). For example, the median (across subjects and

grade levels) standard deviation of mean percentile rank across districts was 6.5, while the median standard deviation of mean residual ranks across districts was 9.8. With less variation to explain, the imputation procedure is more accurate for conditional status norms.

A primary limitation of the study is that the results are specific to our special case of estimating nationally-referenced norms for the ACT Aspire summative assessment system. The sample size, pattern of intermittent missing data across subject areas and grade levels, and participation rates among states and districts across the United States are all unique to our special case, and all of these factors presumably affect the performance of the imputation procedure. This study does not inform the conditions for which the imputation-based estimation method works best. Additional research is needed to study the performance of the procedure under different scenarios of sample size, intermittent missingness, and selection factors for state and district participation in the assessment system. A simulation study could be conducted by simulating test scores for the entire population, removing data through specified missing data processes, and then using the imputation-based approach to try to estimate the population distributions.

While we showed that the imputation-based procedure worked well for estimating nationally-referenced conditional status norms, we only examined the case where one year of prior test scores are used. Conditional status models such as the SGP model can accommodate multiple years of test scores, and in general the reliability of SGPs increase with the inclusion of more prior year test scores. It's possible that the performance of the imputation-based procedure for estimating conditional status norms is affected by the number of prior year test scores included.

Analytic approaches other than imputation could have been used to estimate nationally-referenced norms. We also considered methods based on propensity scores (Rosenbaum & Rubin, 1983). For example, the probability of participating in the assessment system could have been modeled using the same variables used for the imputation model (e.g., student demographics, school variables, and measures of district mean achievement). Inverse probability of treatment weighting (Rosenbaum, 1987) could then be used to assign weights to members of the sample so that the sample is representative of the population. One advantage of the propensity score approach is that we need not assume that the test scores are MVN, and the resulting population frequency distributions need not be normal. A hybrid approach that combines imputation with propensity score weighting is also possible. For example, imputation could be used only for intermittently missing data, and then the propensity score model could be applied. Additional research is needed to examine this method in comparison to the imputation-based approach.

References

ACT. (2017). *ACT Aspire Summative Technical Manual, 2017 Version 4*. Iowa City, IA: ACT.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Betebenner, D. W. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*(4), 42–51.

Betebenner, D. W., VanIwaarden, A., Domingue, B., & Shang, Y. (2017). SGP: Student growth percentiles & percentile growth trajectories (R package version 1.7-0.0) [computer software]. URL: sgp.io

Braun, H., & Qian, J. (2008). *Mapping state standards to the NAEP scale* (ETS Research Report RR-08-57). Princeton, NJ: Educational Testing Service.

Castellano, K. E., & Ho, A. D. (2013). *A practitioner's guide to growth models*. Washington, DC: Council of Chief State School Officers.

Castellano, K. E. & Ho, A. D. (2015). Practical differences among aggregate-level conditional status metrics: From median student growth percentiles to value-added models. *Journal of Educational and Behavior Statistics, 40*(1), 35-68.

Glander, M. (2016). *Documentation to the 2014–15 Common Core of Data (CCD) Universe Files* (NCES 2016-077). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved from https://nces.ed.gov/ccd/pdf/2016077_Documentation_062916.pdf.

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

Little, R. J. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, *6*(3), 287-296.

National Center for Education Statistics. (2009). *The nation's report card: An overview of procedures for the NAEP Assessment* (NCES 2009–493) U.S. Department of Education. Institute of Education Sciences. National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Phillips, G. W. (2007). *Expressing international education achievement in terms of U.S.* Retrieved from https://eric.ed.gov/?id=ED496205.

performance standards: Linking NAEP achievement levels to TIMSS. Washington, DC: American Institutes for Research.

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

Reardon, S. F., Kalogrides, D., & Ho, A. (2017). *Linking U.S. school district test score distributions to a common scale* (CEPA Working Paper No.16-09). Stanford, CA: Stanford Center for Education Policy Analysis. Retrieved from http://cepa.stanford.edu/wp16-09.

Reardon, S. F., Shear, B.R., Castellano, K.E., & Ho, A. (2017). Using heteroskedastic ordered probit models to recover moments of continuous test score distributions from coarsened data. *Journal of Educational and Behavioral Statistics, 42*(1), 3-45.

Rosenbaum, P. R. (1987). Model-based direct adjustment. *The Journal of the American Statistical Association, 82*(398), 387–394.

Rosenbaum, P. R., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.

Rubin, D. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592.

SAS Institute Inc. (2011). *SAS/STAT® 9.3 user's guide*. Cary, NC: SAS Institute Inc.

The Nation's Report Card. (2017). *NAEP data explorer*. Retrieved from https://www.nationsreportcard.gov/ndecore/landing

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multiple imputation by chained equations in R. *Journal of Statistical Software, 45*(3).

Yuan, Y. (2011). Multiple imputation using SAS Software. *Journal of Statistical Software, 45*(6), 1-25.