The Implications of Population Changes on Generalization and Study Design

Forthcoming in the *Journal of Research on Educational Effectiveness*

1. Wendy Chan, Ph.D. (Corresponding Author)
   Graduate School of Education, University of Pennsylvania
   Email: wechan@upenn.edu
   ORCiD: https://orcid.org/0000-0002-0933-9532

2. Jimin Oh
   Graduate School of Education, University of Pennsylvania
   Email: ohjimin@upenn.edu

3. Peihao Luo
   Graduate School of Education, University of Pennsylvania
   Email: peihao@upenn.edu

Abstract

Findings from experimental studies have increasingly been used to inform policy in school settings. Thus far, the populations in many of these studies are typically defined in a cross-sectional context; namely, the populations are defined in the same academic year in which the study took place or the population is defined at a fixed time point. This study assesses the extent to which the composition (observable characteristics) of a population changes over time and the implications of these changes for generalization and study design. We center our discussion around two case studies, both of which were cluster randomized trials in education. For each study, we collected multiple years of population level data and analyzed the types of changes that occurred in each study population over the given time period. Overall, we find that the most consistent changes are associated with the racial and demographic composition of the student populations. We discuss the implications of these changes and how they affect the types of populations for whom the findings of a study will be relevant.

Keywords: longitudinal, population, generalizability, propensity scores, study design

Experimental studies have become increasingly common as a tool used to evaluate the causal impact of treatments or interventions. Over the past decade, with the steady rise in these studies, policymakers have grown interested in understanding the extent to which study results apply or *generalize* to target populations of inference. However, the strongest tool for generalization is random or probability sampling, which is rarely used in practice, particularly in educational studies (Olsen, Bell, Orr, & Stuart, 2013). As a result, statisticians have developed methods to improve the generalizability of findings from nonrandom samples using propensity scores (Cole & Stuart, 2010; Stuart, Cole, Bradshaw, & Leaf, 2011; Tipton, 2013; O'Muircheartaigh & Hedges, 2014). These include the use of matching (Stuart, 2010) and reweighting estimators (see, for e.g., Kern, Stuart, Hill, & Green, 2016) that adjust treatment effect estimates based on characteristics of the population of inference.

Thus far, the populations in many evaluation studies (and many generalization studies) have typically been specified in a cross-sectional context; namely, the populations are defined in the same (academic) year in which the study took place, or there is a single population defined at a fixed time point. However, populations of students and the composition of schools in a population can change over time, leaving open the question of how these changes potentially affect the generalizability and design of current evaluation studies. This is an important consideration for two reasons. First, if the composition (observable characteristics) of a population changes, this affects generalizations of treatment effects since individuals in the sample may not necessarily be comparable to individuals in the target population. As a result, treatment effect estimates may be less useful, whereby "useful," we refer to a treatment effect estimate that is both precise and unbiased for the target population. Second, if populations of inference do change over time, this has implications for the design of future studies. For

example, suppose a study was conducted in a school district to evaluate the effectiveness of a reading program on improving comprehension and literacy among third grade students. Suppose the proportion of English Language Learners (ELLs) or the proportion of students who qualified for free and reduced-priced lunch increased in the following years. In this case, researchers may want to design their study sample to reflect future student demographics. This would ensure that the recruited sample is representative of the target populations of students and that the program impacts are generalizable to these students.

The purpose of this article is to provide the first empirical evidence on the extent to which the compositions of populations change over time. We address this through a discussion of two case studies, both of which were cluster randomized trials in education, and which were also analyzed in prior generalization studies. For each empirical example, we describe the population changes based on a set of observable school covariates (characteristics) that collectively provide information on student and teacher demographics. We frame our discussion of population changes around four main groups: a "broad" population (such as all public schools in the state) and three different subpopulations of schools with shared characteristics (such as all urban schools). The goal of the subpopulation comparison is to determine whether the patterns of changes observed in the full population are also seen in subgroups within the broader population. Using the comparisons from the two case studies, we then discuss the implications of population changes for two main areas of evaluation research, generalization and study design, and highlight several points of consideration when assessing the effects of longitudinal changes in a population on future evaluation research.

The article is organized as follows. First, we state the three main research questions and describe the motivation of each in the context of the study. Then, we introduce the case studies,

describe the population changes that took place and the similarities and differences in the trends between the broad populations and subpopulations. We discuss the implications of population changes for generalization and study design. For generalization, we center our discussion around two statistics, the generalizability index and overlap measure, used to quantify the compositional similarity between the sample and the population. Finally, we discuss the overall implications of longitudinal changes in a target population on current and future evaluation research.

Research Questions

Our study is motivated by the following three research questions:

1) To what extent does the composition of a population, as measured by a set of observable characteristics, change over multiple years and what is the "pace" of these changes?
2) Do similar changes take place within the subpopulations? If so, are the magnitudes of the changes similar between the full and subpopulations?
3) What are the implications of population changes for generalization and for study design?

Question 1 is the general motivating question of the analyses of the case studies. In addition to examining compositional changes in a population, we also assess how "fast" these changes occur. For example, if the proportion of English Language Learners (ELLs) among schools in a state increases on average, does this increase happen within two years or does it take place gradually over five years? The pace at which changes in the proportion of ELLs or other variables occur can have implications for the development of local school policies. If, for example, the proportion of ELLs significantly increases within two years, school administrators may consider changes to resource allocations to support these students. Additionally, with respect to generalization, the findings from studies on ELLs may be particularly relevant and timely as the population of ELLs increases. Question 2 is motivated by an interest in whether the

compositional changes that take place at the broad population level also occur within subgroups of the larger population. For example, if the proportion of ELLs increase in the broad population, an important question is whether we see similar increases among the population of urban or rural schools. Furthermore, Question 2 explores the extent to which the magnitudes of changes seen in the full population are similar or different compared to the subpopulations. If, for example, the proportion of ELLs increases by 2 percentage points in the full population each year, but the proportion increases by 5 percentage points among urban schools, this may be a significant difference to local policymakers when deciding upon best policies to support their students. Finally, Question 3 connects the results from the previous research questions to implications for two primary areas of research: generalizability and study design. The generalizability of treatment effects refers to the extent to which the study results from a (nonrandomly selected) sample are applicable to students in a specified inference population. In practice, the strength of generalizations relies on the compositional similarity between the given sample and inference population. If the composition of the population changes within a short period of time, this has implications for the usefulness of treatment effect estimates. Furthermore, these same implications would also affect the design of future studies as compositional changes in the population may impact sampling decisions.

Case Studies

To address the main research questions, we focus on two case studies, SimCalc and Indiana, both cluster randomized trials (CRTs) in education. These case studies were selected for two reasons. First, both examples were analyzed in prior generalization studies (Tipton, 2013; O'Muircheartaigh & Hedges, 2014; Chan, 2017, 2018; Tipton et al., 2017). Because the implications of population changes on generalization is a focus of this study, it is helpful to center the discussion around earlier generalization research. Additionally, the study samples in

both studies were nonrandomly selected from their respective populations of inference, which is common among experimental studies in education (Olsen et al., 2013). As a result, our discussions of population changes and the implications for generalization are potentially relevant for similar studies in other contexts. Second, because SimCalc and Indiana took place in different states (Texas and Indiana, respectively), this allows us to examine the extent to which the trends in the population changes depend on the context of the study. While the case studies do not incorporate all possible contexts, the results are useful in providing preliminary empirical evidence on the extent to which populations change over time and whether these changes are similar (or different) across studies.

For both studies, we collected multiple years of population level data on students to focus on the longitudinal changes in the "same" inference population. To be consistent with prior work, we used the same covariates from the earlier generalization studies on SimCalc and Indiana. The goal is to observe trends in the covariate distributions for the population across a given time period. We specified the preliminary inference population for each study as all schools meeting a set of eligibility criteria in the entire state (Texas for SimCalc and Indiana for the Indiana CRT). For SimCalc, this included all public schools that served 7^{th} grade students in Texas. For Indiana, this included all public K – 8 (elementary to middle) schools. However, in the original generalization studies for Indiana, Chan (2017) and Tipton et al. (2017) excluded a subset of schools that were considered different from the study sample. In their analyses, schools with over 95% male students, schools with over 95% ELLs, among others, were dropped. To be consistent with these studies, we made the same exclusions in defining the full inference population for Indiana in our study.

In addition to the population of schools in the state, we examined trends in the same covariates among three subpopulations: urban schools, suburban schools, and Title I schools. Our choice of subpopulations was motivated by the types of student groups that are potentially of interest to educational researchers and policymakers. However, different subpopulations can be defined based on other characteristics such as school size and proportion of ELLs. The purpose of the subpopulation analyses is to determine whether there are differences in compositional changes among subgroups of the broader population and if so, the implications of these differences for the generalizability of study findings and policy.

Case Study 1: SimCalc

SimCalc is a mathematics software program that uses computer animations to teach concepts of rate and proportions. A mission of the SimCalc Project, based at the James J. Kaput Center at the University of Massachusetts, Dartmouth, is to provide students in disadvantaged environments with opportunities to learn advanced mathematics (Kaput, 1997). In addition to the software, the SimCalc Project also provided professional development workshops for teachers to strengthen their mathematical content knowledge, to learn to use the curriculum materials associated with SimCalc, and to specifically plan for the use of the materials (Roschelle et al., 2010, p. 847). To assess the impact of SimCalc on mathematics achievement among 7^{th} graders, the research firm SRI International implemented two cluster randomized experiments, one of which was a pilot study, on a combined sample of 92 middle schools in Texas. These studies took place during the 2008 – 2009 academic year. Among the 92 schools, 45 were randomly assigned to receive SimCalc and the remaining 47 were assigned to control. In the original study, the principal investigators found a statistically significant main effect of 1.438 ($SE = 0.143$, $p < 0.001$) in students' mathematics gain scores, implying that students in schools using SimCalc experienced

larger gains than students in control schools (Roschelle et al., 2010). The scores were based on the Texas Assessment of Knowledge and Skills (TAKS) and the treatment effect is standardized in relation to the between-school variance.

Although every effort was made to select a random sample of schools, the SimCalc sample was not a probability sample of Texas middle schools. Schools were primarily recruited through the Charles A. Dana Center at the University of Texas and regional education service centers throughout Texas (Roschelle et al., 2010, p.855). Tipton (2013) and O'Muircheartaigh and Hedges (2014) assessed the generalizability of the SimCalc study to the population of Texas schools using a subclassification estimator based on propensity scores. The population of inference in both studies comprised 1,713 schools that served 7^{th} grade students in the 2008 – 2009 academic year. Propensity score methods were used to assess the generalizability of the SimCalc and Texas population schools based on 26 covariates deemed relevant in predicting sample selection and the treatment effect. These covariates were taken from the Texas Academic Excellence Indicator System (AEIS). Both O'Muircheartaigh & Hedges (2014) and Tipton (2013) estimated an average treatment effect of 1.452 ($SE = 0.195$, $p < 0.01$) in the original population and Tipton (2013) estimated an effect of 1.430 ($SE = 0.188$, $p < 0.01$) in a subpopulation of schools that had the greatest similarity in covariate distributions to the SimCalc sample. These estimates imply that a typical population school using SimCalc would experience an overall average gain in students' scores.

The SimCalc study was based on the population of Texas schools defined in 2008 – 2009. An important question is how changes in the composition of the population would affect the generalizability of the estimates and the design of future evaluation studies in the same population. To assess the compositional changes among Texas schools, we collected 9 years of

data from 2008 – 2009 (SimCalc study year) to 2016 – 2017 (most recent year available) using information from AEIS and the Texas Academic Performance Report (TAPR)[1]. Following Tipton (2013) and O'Muircheartaigh & Hedges (2014), the target populations in each year comprised public schools that served 7th grade students. We attempted, to the extent possible, to collect information on the same 26 covariates as in the original studies. However, due to changes in the reporting system, we were unable to secure all the original variables and had to use several alternative covariates instead. Table 1 gives the final covariate set used in the analyses and their summary statistics. The last two rows correspond to covariates in the original generalization studies that were not part of the final analyses for this study. The covariates in bold are additional variables that were not in the original generalization studies, but which were included in our analyses.

TABLE 1

Data Challenges

In the process of creating the final analytic data file, we encountered several data challenges. First, because of inconsistencies and missing values in the data, the final dataset (across all years) comprised 936 population schools and 63 SimCalc study schools for a combined sample of 999 schools. Second, the Texas Education Agency (TEA) adopted a new statewide standardized testing system, STAAR (State of Texas Assessment of Academic Readiness), in 2012 – 2013, which replaced the prior TAKS (Texas Assessment of Knowledge and Skills) system. As a result, variables that captured assessment data such as the % of students with commended performance in the original studies and the "G3-G11" variables used in our analyses were not necessarily comparable across this period since they depended on the system that was

[1] Note that AEIS switched to TAPR in the 2012 – 2013 academic year.

in place in the given academic year. Despite several attempts to clarify the alignment between STAAR and TAKS with TEA representatives, there was no conclusive statement on whether the standards in each system were comparable. Thus, while the assessment variables are included, it is important to note that the changes in the testing system affect the comparability of the values.

Changes to Definitions of Variables

In addition to changes in the testing system, the TEA also modified the definitions of multiple variables. The main changes to variable definitions were associated with DAEP (students in disciplinary alternative education programs) and the % of students at risk. DAEP classrooms are alternative educational settings for students who are removed from their regular classes for mandatory or discretionary disciplinary reasons (TEA, 2007). For DAEP, the TEA included 19 "reason codes" to determine placement in the alternative education programs (TEA, 2007). Reason codes refer to specific disciplinary actions taken by the school or district to address concerns related to students' behavior and conduct.[2] In contrast to prior years, in 2009 – 2010, the TEA included four additional codes (23 codes) to determine placement in DAEP and this change resulted in significantly different values for this variable (see Table 1). In addition to DAEP, the TEA also modified the definition of the % of students at risk variable. Prior to 2013 – 2014, the % of students at risk only included students under 21 years of age. However, in the years following 2014, the age limit was extended to 26 years, which likely affected the year-to-year changes in this variable. As a final note, though the changes in testing system potentially affected the reclassification of ELLs, we found no empirical evidence supporting this claim. However, Table 1 illustrates that there was a sharp increase in the % of ELLs (by 1.13 percentage points) in 2012 – 2013, coinciding with the year when the TEA first used the STAAR

[2] For example, code # 5 refers to an out of school suspension (http://ritter.tea.state.tx.us/peims/standards/1314/c164.html).

assessment system. Although the empirical evidence is inconclusive, in the following sections, we note that the changes associated with the ELLs may be interpreted with caution. A full summary of all changes to variable definitions is given in the Appendix.

SimCalc: Changes in the Full Population

In this section, we begin by describing the changes in the full population. Figure 1 provides plots of the individual covariate changes in SimCalc for both the full population and the three subpopulations of urban, suburban, and Title I schools. Figure 1 illustrates that, over the 9-year period, the most consistent changes among Texas schools (full population) are associated with the racial and demographic composition of the student and teacher populations. First, the percentage of Hispanic students and teachers consistently increased from year to year, a trend that was noted in a recent TEA report (TEA, 2018). The largest growth in the Hispanic student population takes place in 2010 – 2011, two years after SimCalc, where the percentage of Hispanic students increases by 1.81 percentage points in a single year. Coincidentally, the same year is associated with the largest growth in the population of Hispanic teachers (1.13 percentage points). Second, although the trends are less consistent, the average percentage of ELLs also increases over this time period, rising by about 1.2 percentage points over two consecutive years between 2012 – 2013 and 2013 – 2014. Although this change may be interpreted with caution (see previous section on changes to the definition of variables), there was a slight increasing trend in the percentage of ELLs prior to 2012 – 2013. In addition to the growth of the ELL population, the percentage of students who qualified for free- and reduced-priced lunch (FRPL) and the percentage of students at risk increased overall. The largest growth in FRPL students (2.5 percentage points) takes place in 2009 – 2010, the year immediately following SimCalc.

FIGURE 1

The trends in Figure 1 also suggest that there were changes to school climates. Many Texas schools, on average, employed teachers with fewer years of teaching experience, which is implied by the overall drop in full-time teachers with over 20 years of experience. Among students, the average retention and mobility rate decreased on average and some of the larger drops in student retention are seen in the years immediately following SimCalc. Additionally, there was a consistent drop in the percentage of students who were in disciplinary alternative education programs (DAEP). Note that the unusual jump in DAEP between 2008 – 2009 and 2009 – 2010 was due to the TEA's inclusion of additional reason codes, and as a result, we consider the average DAEP in the latter year as a different variable altogether. Our observations for DAEP are thus primarily based on the remaining years.

Overall, with the increases in percentages of ELLs and students at risk, the trends in SimCalc suggest that many Texas schools experienced a growth in student populations with greater academic needs. While the definition of the percentage of students at risk changed in 2013 – 2014, there was a consistent increasing trend in the years following this period. Although the magnitude of the population changes was small (all less than five percentage points), the changes were positive (in direction) overall and for some covariates, the changes were consistently positive. These findings have important policy implications since they affect the decisions that school leaders and instructors make in determining the appropriate resource allocations to support the students. With the growth in Hispanic students, schools likely prioritized the availability of bilingual resources for both students and parents. In addition to the growth in several student populations, the percentage of Hispanic teachers also increased, which contributed to the racial diversity of the teacher populations in many schools. Importantly, the

growth in the population of Hispanic teachers likely provided a key resource and source of support for Hispanic students.

Over the same 9-year period, Texas population schools experienced changes in school climate and culture. The decline in the number of DAEP students and in the retention and mobility rates among students likely had positive effects on school culture and the academic experiences of students. Additionally, many schools hired more teachers with fewer years of experience, which potentially introduced new forms of pedagogy and teaching resources to school settings. However, it is possible that the increase in the percentage of new teachers implied more instability in school environments since new teachers, on average, experience higher turnover and mobility rates. Although our discussion of the compositional changes only speculates on their impacts for students and teachers, it highlights possible implications of these changes, which are important considerations for researchers.

SimCalc: Changes in the Subpopulations

Given the changes at the broad population level, an important question is whether similar changes occur at the subpopulation level. Overall, as shown in Figure 1, the three subpopulations experienced similar shifts in racial composition and school climate as the full population. The percentage of Hispanic students, ELLs, and students at risk increased and the percentage of DAEP students dropped, on average. However, while the trends were similar among the subgroups, the specific demographic and racial composition of the schools were different among the subpopulations. For example, urban schools consistently had the largest percentage of Hispanic students compared to suburban schools and schools in the full population each year. While the proportion of Hispanic students was increasing across all populations, the implications of this growth were likely different for urban schools which already had large existing

populations of Hispanic students. As a result, the impact of a growing Hispanic population was potentially smaller for urban schools (compared to suburban schools) as resources for these students were likely already in place.

Perhaps the most notable differences between the full population and subpopulations were in the percentage of African American students and teachers. Among suburban schools, the percentage of African American students increased on average, while this percentage decreased each year among Title I schools. The percentage of African American teachers decreased overall among the three subpopulations, but like the proportion of Hispanic students, there were differences in the magnitudes of this change. The percentage of African American teachers dropped by about 0.4 percentage points among urban schools in the year immediately following SimCalc. In comparison, the same percentage dropped by about 0.2 percentage points among suburban schools and schools in the full population. Although the magnitude of these changes are small (like in the full population), the differences in percentage change of African American students and teachers imply that some subgroups of schools experienced more pronounced shifts in the racial diversity of their school populations. Furthermore, these differences imply that policy decisions, such as changes to accountability standards and budget cuts, likely affected subpopulation schools in different ways.

Case Study 2: Indiana

Our second case study is based on a CRT that took place in Indiana from 2009 – 2010. During this year, the Indiana Department of Education and the Indiana State Board of Education implemented a new assessment system designed to measure annual student growth and to provide feedback to teachers (Konstantopoulos, Miller, & van der Ploeg, 2013). Fifty-four K – 8 (elementary to middle) schools from the state of Indiana volunteered to participate in a CRT to

evaluate the effect of the assessment system on academic achievement. Of the sample of 54 schools, approximately half were randomly assigned to use the state's new assessment system (treatment) while the remaining schools were assigned to control. In the treatment schools, students were given four diagnostic assessments that were aligned with the Indiana state test, and their teachers received online reports of their performance to dynamically guide their instruction in the periods leading up to the state exam. The treatment effect of the assessment system was measured using the Indiana Statewide Testing for Educational Progress-Plus (ISTEP+) scores in English Language Arts (ELA) and mathematics.

Chan (2017) and Tipton et al. (2017) assessed the generalizability of the Indiana study to the target population of 1,460 K – 8 schools in Indiana during the 2009 – 2010 academic year. Both studies used matching methods to compare the study and Indiana population schools on 14 covariates, which included continuous measures such as pretest scores, school size, attendance, and binary measures such as Title I status. In the original study (Konstantopoulos et al., 2013) and in both generalization studies, the effect of the benchmark assessment system was not statistically significant.

To assess the extent to which the population of Indiana schools changed over time, we collected 7 years of population level data from 2009 – 2010 (Indiana CRT study year) to 2015 – 2016 (most recent year available) using information from the Common Core of Data (CCD; https://nces.ed.gov/ccd/), the Indiana Department of Education (IDOE), and the United States Census. For each academic year, the target population of inference was specified as all public elementary and secondary schools. We excluded the same types of schools that were removed in the prior generalization studies (schools with over 95% male students, schools with over 95% ELLs, etc). We successfully secured most of the original covariates and collected additional

information on urbanicity and the proportion of students who met a "Pass" criterion for the ISTEP+ assessments. Our analyses are based on a final covariate set of 16 variables. Table 2 provides the summary statistics for the covariates over the 7-year period. The last three rows of the table refer to covariates that were in the original generalization studies but were not included in our analyses. The covariates in bold are three additional variables that were not in the original studies, but which were included in our analyses. Because of missing values for some covariates, our final data analytic file comprised 1,225 population schools and 41 Indiana CRT schools across the years.

TABLE 2

Changes to Definitions of Variables

Prior to our discussion of the population changes, we used the IDOE archives to determine whether the definitions of the 16 covariates in Table 2 changed over the given 7-year period. Overall, most of the variable definitions did not change and the only changes were associated with the ELA and Math pass ratio. Beginning in 2014 – 2015, the IDOE transitioned from the Indiana Academic Standards to the college- and career-ready Indiana Academic Standards, which included changes to several state assessments for elementary and middle school students.[3] As a result, the ELA and math pass ratios in the years prior to 2014 – 2015 and afterwards were likely measured differently. While these changes affect the comparability of the populations, our analysis of the Indiana study largely focuses on the period before 2014 – 2015. Nevertheless, with respect to the changes in the ELA and math pass ratios, we note that the comparisons before and after 2014 should likely be interpreted with caution given the changes to the academic standards.

[3] https://www.doe.in.gov/sites/default/files/standards/assessmentandaccountabilityaug2014.pdf

Indiana: Changes in the Full Population

Figure 2 provides plots of the individual covariate changes for the Indiana study for both the full population and subpopulations of urban, suburban, and Title I schools. Interestingly, in contrast to SimCalc, many covariates in the Indiana populations largely remained the same over the 7-year time period. The proportion of urban schools, for example, only increased in the two years between 2012 – 2013 and 2015 – 2016, while remaining unchanged for the other years. However, there were several important exceptions to this trend. First, the percentage of white students declined overall, suggesting similar shifts in racial composition like those observed in SimCalc. Second, the proportion of students eligible for FRPL increased on average, which coincides with an overall increase in the proportion of schools with schoolwide Title I status. For the latter, the largest increase (9 percentage points) takes place in 2010 – 2011, the year immediately following the Indiana CRT. Third, the percentage of full-time teachers decreased on average while the pupil teacher ratio increased. The largest increase in the pupil teacher ratio (13%) occurred in 2014 – 2015. This suggests that the demographic composition of the teacher population changed and that students likely experienced larger class sizes over the years.[4] As a final important trend, the average pass ratio in ELA and Math improved, with the exception of the final year in 2015 – 2016, where the average pass ratio dropped by nearly 30% compared to the previous year. However, this is likely due to the changes in the IDOE academic standards.

FIGURE 2

[4] We investigated this trend further and found that the changes were largely concentrated in one county, St. Joseph County, which enacted a 3% budget cut in 2014 and a 5% budget cut in 2015. These cuts resulted in teacher layoffs. Additionally, Indiana enacted multiple cuts to retirement benefits for teachers, which prompted many teachers to retire early or to leave the state. For additional details, see https://wsbt.com/news/local/st-joseph-county-enacts-budget-cuts-hiring-freeze-layoffs-through-attrition. We assessed the trends without the schools in St. Joseph county. Without this county, the pupil teacher ratio increased by 4% in the full population.

Overall, Figure 2 illustrates that while many covariates did not change much among the population of Indiana schools, the few notable changes were associated with student demographics and features of the school environments. The decreases in the proportion of white students, combined with the increase in FRPL students and in the proportion of Title I schools, imply that in many Indiana schools, there was a growth in student populations that were diverse, both racially and in the types of academic needs. Like SimCalc, this potentially had policy implications as school leaders considered the kinds of resources needed to support these students. In addition to changes in the demographic composition of students, the average class size increased over the years, which likely impacted various aspects of the school environment, including instruction and allocation of academic resources. Interestingly, these changes did not appear to impact schools' performance as the pass ratios in ELA and math largely improved over the 7-year time period.

Indiana: Changes in the Subpopulations

Figure 2 also provides the plots of covariate changes for the subpopulations. Across the three subpopulations, we see similar trends in the percentage of full-time teachers and the pupil teacher ratio; namely, the percentage of full-time teachers declined overall, and the pupil teacher ratio increased on average. Additionally, the average ELA and math pass ratio improved over the years for all subpopulations, and the size of the improvements is similar to those in the full population. However, the compositional changes in the subgroups differed from the full population in several ways. First, among urban and Title I schools, the proportion of ELLs increased on average while the trends among the suburban schools and in the full population were mixed. Second, while the ratio of students in special education classes appear to be decreasing on average among urban schools, this ratio is mainly increasing in suburban and Title

I schools. Importantly, while the ratio of white students was decreasing on average in the full population, this ratio *consistently* decreased over all 7 years for the three subpopulations. For some subpopulations, such as urban schools, the ratio of white students decreased by 2% almost every year.

While the overall trends were similar in both the full population and the subpopulations, the magnitudes of the changes were different. Among urban schools, for example, the pupil teacher ratio increased on average, but from 2014 – 2015, this ratio jumped by 46% compared to the 13% and 8% observed in the full population and the subpopulation of suburban schools, respectively. As a result, students in many urban schools experienced dramatic increases in class size compared to other schools in the state. Furthermore, the overall increase in the proportion of ELLs imply that in addition to larger classes, many urban schools experienced a growth in student populations that were both diverse and that likely required additional academic support. Furthermore, like SimCalc, the decrease in the ratio of white students among all subpopulations signify that the student populations in many schools became more racially diverse with some schools (urban) experiencing more significant shifts than others.

Implications of Changing Population Demographics

Both SimCalc and Indiana illustrated that many population schools experienced notable changes in the demographic composition of its students and teachers. Despite some differences between the two examples, a common trend in both studies was that there was greater racial diversity among the student populations. An important question is whether these compositional changes in the target population have implications for current practices in evaluation studies, particularly in longitudinal studies. In this section, we focus on the implications for two areas of evaluation research, generalizability and the design and planning of future studies, and discuss several

important considerations for researchers in the process of understanding the effects of longitudinal changes in the population.

Implications for Generalizability

Generalization research thus far has primarily focused on the extent to which findings from experimental studies apply to individuals in a target population of inference. Much of the recent work in this area has centered on the development and application of statistical methods to improve generalizations, particularly from studies with nonrandomly selected samples, as seen in the case studies. Because estimates of population average treatment effects (PATEs) are generally biased in nonrandomly selected study samples, propensity score methods have been proposed to address the bias due to potentially systematic differences between schools that select and do not select to participate in a study (Shadish et al., 2002; Cole & Stuart, 2010; Stuart et al., 2011; Tipton, 2013; O'Muircheartaigh & Hedges, 2014). Since their development, propensity score methods have been used to improve generalizations by matching or reweighting the study sample to be compositionally "like" schools in the population of inference (Stuart, 2010). If the assumptions for propensity score methods are met, researchers can derive bias-reduced estimates of the PATE.

A key consideration in propensity score methods for generalization is that the schools in the sample and population must be compositionally "similar." This similarity is based on a set of observable characteristics that are assumed to affect sample selection and to completely explain the variation in treatment effects between the sample and population, an assumption known as *sampling ignorability* (Tipton, 2013). As the results of our case studies illustrate, the composition of the same inference population can change over time. This motivates the question of how these

changes affect the generalizability of the study sample to the target inference population and more importantly, how the changes affect the generalizability of the study's results.

Measures for Assessing Generalizability

To assess the relationship between population changes and the generalizability of the study samples, we turn to two statistical measures to assess generalizability, both based on propensity scores. In this section, we give a brief overview of propensity scores and formally define the statistical measures. Given a population *P* of *N* schools, suppose a sample *S* of *n* schools are nonrandomly selected to participate in a study. Throughout, we assume that the inference population has been well-specified and defined. Let *Z* denote a sample selection indicator such that *Z = 1* if the school is in the experimental sample *S* and *Z = 0* otherwise. For each school in *P,* let **X** denote a vector of observable covariates, which may include a combination of continuous variables such as academic assessment scores and categorial variables such as urbanicity. The sampling propensity score is defined as:

$$s(\boldsymbol{X}) = \Pr(Z = 1|\boldsymbol{X}) \tag{1}$$

The sampling propensity score is the conditional probability of selection into the sample, conditional on the observable covariates **X**. The estimated propensity scores are used to match schools in the sample with those in the population so that the resulting groups are compositionally similar. Notably, propensity scores are balancing scores where schools with similar propensity scores have similar distributions in the covariates of the propensity score model (Rosenbaum & Rubin, 1983). A common method of estimating $s(\boldsymbol{X})$ is with logistic regression based on *p* covariates:

$$\text{logit}\big(s(\boldsymbol{X})\big) = \log(s(\boldsymbol{X})/1 - s(\boldsymbol{X})) \ = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_p X_p \tag{2}$$

The validity of propensity score methods for generalization relies on four main assumptions (Stuart et al., 2011; Tipton, 2013). Among these, sampling ignorability requires that the propensity score model includes *all* covariates **X** that affect both treatment effect heterogeneity and sample selection. Additionally, sampling ignorability requires that every school in the population (sample) has a comparable school in the sample (population) and no school has $s(\boldsymbol{X}) \approx 0$. If, for example, the Indiana study was comprised of mainly rural schools and students in these schools responded differently to the assessment system compared to students in urban schools, then urbanicity should be included in the propensity score model to satisfy sampling ignorability.

If the assumptions for propensity scores are met, the generalizability of the study sample can be assessed using two statistical measures, the generalizability index (Tipton, 2014) and distributional overlap. These two measures, both based on propensity scores, range from 0 to 1 where values close to 1 are associated with stronger generalizability; that is, the closer the value is to 1, the more compositionally similar the sample is to the population on the given covariates. The generalizability index (referred to as the B-index in Tipton, 2014) is based on the Bhattacharyya coefficient (1943) and uses the density functions of the estimated propensity score logits in (2) to quantify the similarity in propensity score distributions between the sample and the population. Tipton (2014) defined four ranges for the B-index where values in the range $[0.90, 1.00]$ indicate "very high generalizability," values in $[0.80, 0.90)$ indicate "high generalizability," values in $[0.50, 0.80)$ indicate "medium generalizability" and values below 0.50 indicate "low generalizability."

Distributional overlap (referred to as overlap hereafter) is defined by the proportion of schools in the population whose estimated propensity scores fall in the middle 90% range of the sample propensity scores (Chan, 2020). Formally, overlap is defined as

$$\Omega = \int_{q_{min}}^{q_{max}} f_p dp \quad (3)$$

where f_p is the density function of the propensity scores in the population, and q_{min}, q_{max} are the 5th and 95th percentile, respectively, of the propensity scores in the sample. Overlap values close to 1 indicate that a greater proportion of population schools have propensity scores that are in the range of the sample, which implies a stronger degree of similarity between the sample and population. Note that q_{min} and q_{max} can be based on the minimum and maximum propensity score in the sample, but we follow the observational studies literature and define the min and max to be the 5th and 95th percentile of the propensity score distribution in the sample.

Because the B-index and overlap summarize the similarity between individuals in the sample and the population, compositional changes in the population could affect the values of both measures. If, for example, the changes in the population are associated with smaller values of either measure, this implies weaker compositional similarity between schools in the sample and population. Importantly, if the covariates **X** satisfy the assumptions for propensity scores (sampling ignorability), then changes in the values of the B-index and overlap affect the precision and bias of treatment effect estimates. In following sections, we assess the effects of compositional changes on the generalizability of the SimCalc and Indiana study samples by computing values of the B-index and overlap, hereafter referred to as the *generalizability statistics.* For these analyses, we consider the population schools in each academic year as a separate inference population and use the trends in the values of the generalizability statistics to assess the relationship between population changes and the generalizability of the study samples.

Note that the estimated values of the generalizability statistics assume that the definitions of the covariates in the propensity scores are consistent between the sample and target population. Although the definitions of some covariates changed in the time periods in our case studies, the comparisons between the populations are still valid if the differences in definitions are considered small. However, with significantly large differences, comparisons based on the generalizability statistics may be inappropriate.

Exclusion of Covariates for SimCalc

Given the changes to the definitions of several variables in the case studies, an important question is whether these covariates should be included in the generalization analysis. For both case studies, we considered the changes to all variable definitions and decided to only exclude covariates for the SimCalc study; namely, the DAEP and G3-G11 variables. Our decision was motivated by two factors. First, because of changes in their definition, we believed that the values of the DAEP and G3-G11 variables were not necessarily comparable across the 9-year time period. When the TEA changed the approach used to measure DAEP in 2009 – 2010, this led to an unusually large value in the given year. Similarly, for the G3-G11 variables, because the testing system in Texas changed in 2012 – 2013, the assessment variables before and after the change may not be comparable. We believed that including the DAEP and G3-G11 variables would potentially contribute to systematic differences, solely based on differences in definition, between the sample and populations, which would consequently affect the values of the generalizability statistics. A second reason for excluding the DAEP and G3-G11 variables is that their inclusion led to population schools with estimated propensity scores that were close to zero. Population schools with propensity scores that are approximately zero suggest that there is an overlap issue and that these schools may not have a comparable set of sample schools for

generalization. In this case, researchers may choose to exclude the schools whose propensity scores are close to zero to satisfy the sampling ignorability assumption for generalization (Tipton, 2013). Note that if schools are excluded, this redefines the inference population for generalization. Alternatively, researchers can base generalizations on a subset of covariates that do not cause instability in the propensity score estimation. However, the choice of covariates to include in the propensity score model should be motivated by substantive theory on the variables that potentially moderate treatment effects between the sample and population. While researchers may choose to omit covariates to minimize instability in the propensity score estimation, it is important to note that this omission leads to biased results if the excluded covariates are important moderators. In either case, exclusions, whether it is of the schools or covariates, affect inferences from generalizations. In the Appendix, we describe the results of some preliminary analyses that included DAEP and the G3-G11 variables. However, because the instability problem was persistent and because we aimed to minimize the potential bias due to systematic differences from definition changes, we ultimately decided to omit the variables entirely.

Although similar changes to definition were observed in the Indiana CRT academic performance variables, we chose to include all the covariates in the generalizability analysis. Since the changes were confined towards the end of the data period, we believe that including the academic performance variables did not affect the overall trends in the Indiana study analysis. Additionally, the inclusion of the covariates did not cause the same kind of instability in propensity score estimation that occurred in the SimCalc study.

While our decisions to include and exclude certain variables in each study was somewhat informal, we offer two considerations for researchers in deciding when covariates should be removed from the analysis. First, if the definition of the covariate changes to the extent that the

variable before and after the change are not comparable, then we believe that the covariate should be excluded. This can occur with definition changes (such as the DAEP variable in SimCalc), with external system changes (such as the G3-G11 variables in SimCalc), or with changes in the type of instrument or scale used to measure the covariate. In these examples, the changes in measurement and definition can lead to systematic differences among schools, which affects the bias in the parameter estimates. For generalization studies with nonrandom samples, estimates of the PATE are subject to bias from systematic differences due to unmeasured confounders between the sample and population. If changes to the definition of covariates contribute to the systematic differences, then it may be inappropriate to include them in the final analyses. Second, for covariates that have comparatively "minor" changes (such as the enrollment variable in SimCalc), the decision to exclude these variables will depend on whether the differences in measurement would significantly contribute to differences between the sample and population (which would affect the bias of the PATE estimates). We believe that a general consideration is whether the covariates in question would be needed to satisfy the assumptions for propensity scores; namely, whether the covariates moderate both sample selection and treatment effect heterogeneity. If the covariates are important moderators, then including them in the propensity score analysis contributes to bias reduction, which may potentially offset any bias due to differences in measurement and definition. If the covariates do not moderate selection and treatment effect variation, then the researcher may choose to exclude them, particularly if differences in measurement and definition affect the estimated propensity scores. However, this is a difficult assessment and the decision to exclude covariates will ultimately depend on the researchers' judgement of the impact on the bias in the estimates. Regardless, for any covariate

exclusions, we believe it is important to provide justification for the decision and a discussion of the implications it may have for inference.

Generalizability Statistics for Case Studies

Using the covariates in Tables 1 and 2 (with the noted exclusions), Figures 3 and 4 provide the B-index values and overlap for SimCalc, respectively, across each academic year. Figures 5 and 6 provide the associated B-index and overlap values for Indiana. The values of these generalizability statistics quantify the compositional similarity between the original study sample and the inference populations defined by the given academic year. The figures for the B-index are shaded in grey to reflect the four ranges of generalizability in Tipton (2014). In addition to the full population, values of the generalizability statistics are also provided for each of the three subpopulations (denoted by the different lines) to assess the compositional similarity between the original sample and subgroups of the population.

FIGURES 3 - 6

Not surprisingly, in both studies, the demographic and racial compositional changes in the populations are associated with *smaller* values for both generalizability measures. This implies that schools in the study samples and inference population became less compositionally similar on the respective study covariates across the years. For SimCalc, Figures 3 and 4 show that the steepest drop in similarity takes place within the first three years of the study and this is observed in both the full population and three subpopulations. These trends are consistent with the covariate changes seen in Table 1 and Figure 1 where the larger shifts in the racial and demographic composition among population schools took place in the first three years. For Indiana, the generalizability statistics in Figures 5 and 6 illustrate that similar drops in compositional similarity took place within the first two years of the study. For some

subpopulations, such as the suburban schools in SimCalc and the urban schools in Indiana, the values of the generalizability statistics are small to the point that schools in the study and population are considered nearly incomparable (based on the included study covariates) across the years. This is seen in the small values of the B-index (below 0.5) and overlap (below 0.10) for the suburban population in SimCalc (Figures 3 and 4) and the urban population in Indiana (Figures 5 and 6). Between the full population and subpopulations, the study schools are more generalizable (similar) overall to schools in the full population in both studies. This is seen in the larger values of the B-index and overlap for the full population across the years. For Indiana, we conducted an additional analysis with the subpopulation of rural schools and found that the trends among this subgroup of schools were similar to the full population (see Table iii in the Appendix). Interestingly, despite the length of time from the original study, the final B-index values for both studies in the full populations (2016 – 2017 for SimCalc and 2015 – 2016 for Indiana) are in the "middle" range. This implies that there is still some degree of compositional similarity between the original study sample and schools in the inference populations that are defined nearly a decade later.

The trends in the generalizability statistics for SimCalc and Indiana illustrate that the original study samples are the most generalizable (compositionally similar) to the inference populations in the earlier years after the study. This compositional similarity decreases overall for the populations in the later years, but there is still some extent of comparability (on the respective study covariates) between schools in the original study samples and those in the final inference populations. However, despite these findings, we are *not* arguing that the original PATE would still generalize to the inference populations defined later. Generalizing the treatment effect estimates to later populations would require that sampling ignorability holds for

each inference population defined across the years; namely, that the propensity score model includes *all* covariates that moderate treatment effect variation for *each* inference population. Since the covariates that satisfy sampling ignorability for the population in one year may not meet the assumption in a different year, the treatment effect estimates do not necessarily generalize, even for the "same" inference population. For example, consider the SimCalc study and the population of Texas schools in 2008 – 2009 (study year) and 2016 – 2017. To generalize the PATE of SimCalc to the populations in these two years, the propensity scores should include all covariates that moderate treatment effect variation in each year. However, though variables such as the average number of computers per classroom may moderate the treatment effect in the study year (2008 – 2009), this is not necessarily the case for the population in 2016 – 2017, particularly with advancements in technology. In 2016 – 2017, for example, the average number of tablets may be an important moderator of the treatment impact of SimCalc. Thus, while the inference population may be the "same" across a period of time (for example, all public middle schools in Texas each year), the types of variables that affect heterogeneity in treatment effects may be different, especially when we consider generalizations over time.

Figures 3 – 6 highlight another important observation regarding longitudinal changes and generalizations. For both studies, there were instances where the generalizability statistics had larger values in later years compared to earlier years. This suggests that the study sample was *more* compositionally similar (on the observable covariates) to populations defined at a later point in time. This can occur if the differences in observable covariates (used in the propensity scores) between the study sample and the population in the later year were smaller compared to the earlier years. Although these trends suggest that the PATE estimate is potentially more generalizable in a later year (compared to an earlier year), we caution against this conclusion.

While the differences in covariates between the sample and the target population may be smaller in a later year, there may be larger differences in other covariates, including unmeasured ones, over the given time period. Importantly, if these covariates (with larger differences) are needed to satisfy sampling ignorability for the sample and target population, the PATE estimates would be *less* generalizable. Thus, despite the implications that compositional similarity may be stronger in later years, it is important for researchers to consider whether the covariates needed to satisfy sampling ignorability change from year to year and whether the distributions of these covariates differ between the sample and target population. Additionally, note that the variables that satisfy sampling ignorability may include both measured and unmeasured covariates. If the measured covariates that satisfy sampling ignorability change over time, researchers can adjust PATE estimates (perhaps with reweighting) to reflect the population changes. However, this is difficult with unmeasured covariates and researchers should assess the sensitivity of PATE estimates to changes in unmeasured moderators when considering generalizations over time.

Covariates and Sampling Ignorability in Case Studies

Because our discussion thus far has focused on the importance of sampling ignorability, we conducted a final analysis to empirically validate this assumption by assessing the extent to which the observable covariates in each study moderated treatment effect variation and sample selection. For SimCalc and Indiana, we fit a series of linear regression models using the covariates in Tables 1 and 2, respectively, and the outcomes in the original studies. To assess treatment effect heterogeneity, we included two-way interaction terms between the treatment indicator and a subset of covariates (Djebbari & Smith, 2008). We chose the set of covariates for the interaction terms based on prior evidence of treatment effect moderation in the original studies, which included the percentage of FRPL students in SimCalc (Roschelle et al., 2010) and

from our own re-analyses of the data. Note that the original studies for both SimCalc and Indiana used student- and school-level data while our analyses are solely based on school-level data. Using the regression models with the interaction terms, we computed unweighted and weighted estimates of the PATE for each study. The weighted estimates are computed using the inverse propensity scores (Buchanan et al., 2018) based on the logistic regression model in (2) with the respective covariates in Tables 1 and 2. We summarize the findings here and provide the details of the models and the full results in the Appendix (Tables iv – vii).

In both SimCalc and Indiana, we found no evidence of treatment effect moderation. Although covariates such as the % of ELLs and the % of Hispanic students were significant in predicting differences in student gain scores in SimCalc, the insignificant interaction effects with receipt of treatment imply that these variables did not affect treatment heterogeneity. For Indiana, none of the covariates were significant in predicting changes in the outcome and the interaction effects were not significant in both the unweighted and weighted models. Note that the original treatment impact estimate for Indiana was not statistically significant (Konstantopoulos et al., 2013; Tipton et al., 2017). In addition to these models, we fit logistic regression models (results not shown) with the same covariates to determine whether any of the variables were significant in predicting sample selection. For SimCalc, DAEP, the % of ELLs, and the % of beginning teachers were significant in predicting selection into the study. For Indiana, urbanicity was an important predictor for selection. However, while these covariates were significant in affecting sample selection, none of their interactions with the treatment indicator were significant. Thus, the empirical evidence suggests that the covariates in our analyses were not relevant moderators of treatment effect variation and as a result, these variables did not appear to satisfy the sampling ignorability assumption.

Despite the absence of moderator effects, we argue that the population changes illustrated in the case studies still provide important considerations for researchers. Although the covariates in SimCalc and Indiana did not appear to moderate treatment heterogeneity, they may be associated with or "proxies" for variables that are relevant moderators. To give an example, suppose that factors related to school climate and culture are moderators of the treatment impact of the intervention in the SimCalc study. Specifically, suppose that students in schools where the software was actively supported and promoted by school leaders, teachers, and parents experienced different impacts from the intervention. While characteristics related to school climate were not captured in the observable covariates for SimCalc, these factors may be associated with covariates such as the average mathematics class size and the years of experience of the teacher, which were used to estimate the propensity scores. Conceivably, with larger class sizes, teachers may be amenable to using the software to differentiate instruction. Newer teachers may be more open to integrating software in their classrooms and instruction. Although these are hypothetical examples of the connection between the covariates and true moderators, understanding how the covariates (from the administrative data) change may still provide insight into how the true moderators of treatment impacts change. Importantly, if the true moderators include unobservable covariates, this has a serious effect on the bias of the PATE and understanding the extent of population changes may be a useful tool in assessing the validity of generalized treatment impact estimates. If populations experience significant changes in observable covariates, there may be equally significant changes in unobservable covariates, which may affect the appropriateness of generalizations to a target population.

Implications for Study Design

Definition of Inference Population

The types of compositional changes that take place in a population and the pace at which they occur have implications for the design and planning of future evaluation studies. One implication is that the shifts in the racial and demographic composition among schools potentially affect the definition of the inference population of a study. Suppose, for example, that the population of inference was specified as all public schools in a state in a given year. If the racial and demographic composition of the schools change in the following year, the inference populations between the two years may not necessarily be the same (the latter may be more racially diverse). While some studies may specify "broad" populations (such as all schools in a state), our empirical examples illustrate that even broadly defined populations can change, and sometimes within a year. As a result, when researchers design future studies (particularly generalization studies), it will be important to have precise definitions of the inference population and to discuss how these definitions can change, specifically when the findings are disseminated to policymakers and stakeholders.

Implications for Current Studies

In both SimCalc and Indiana, multiple changes occurred that affected the student and teacher populations of schools. Although most of the changes were small in magnitude, for some variables, there was a consistent positive change each year. Given these trends, an important question is whether there are implications for the pace at which current studies in education are designed and conducted. Because various populations experience changes differently, we believe that the implications will depend on the research goals and the target populations of the study. Suppose, for example, that a study was conducted among Texas population schools to evaluate the impact of a professional development program on teacher retention among instructors of racial minority backgrounds. If the proportion of minority teachers decreases over a given time

period, these changes will likely affect the relevance of the study's findings. In particular, if the study was conducted within two years and the proportion of minority teachers sharply dropped in the first year, the findings may be less relevant, particularly among schools with fewer minority instructors. On the other hand, if the proportion of minority teachers largely remained the same across schools or the changes in proportion were small in magnitude, the impact on the relevance of the study's findings would likely be less serious. If compositional changes affect the targeted populations of a study, then design considerations should be made to ensure that the study's findings are still applicable to individuals in the target inference population. This may imply, for example, additional pilot studies to both assess the efficacy of the intervention and changes in the context of the population and study setting. For example, suppose that an investigative team were to conduct one additional pilot study for an intervention. This second pilot study would potentially serve two purposes. One, if the intervention produced a positive and statistically significant impact in the first pilot study, the results of the second pilot can be used to confirm the efficacy of the intervention. Two, if generalization is an area of interest, the time between the first and second pilot study would allow researchers to assess whether there were important changes in the inference population. If the inference population becomes less compositionally similar to the study sample, the investigative team may consider whether generalizations are still appropriate. In this example, additional pilot studies may help researchers determine whether important changes to their study settings and populations occur (depending on the length of the studies) and if so, to incorporate these changes in the discussion of their study findings and their relevance.

Finally, it is important to note that covariates do not necessarily change at the same pace. In the Indiana study, while some covariates such as the percentage of schools with Title I status

increased within a year, other variables such as the ELL ratio largely remained the same over the given time period (in the full population). Additionally, while some covariates, such as the % of Hispanic students in SimCalc consistently increased from year to year, other variables such as the rate of student mobility had less consistent patterns of change. The pace of these changes also affects the relevance of a study's findings, particularly if they concern a specific subpopulation of students. As an example, studies that evaluate the impact of interventions designed to support ELLs will be especially timely among schools where the population of ELLs sharply increases or is consistently increasing from year to year.

Recruitment

Population changes can also inform recruitment and sampling decisions, specifically for generalization studies. When researchers design studies with a focus on the generalizability of the results, it is important to recruit samples that are representative of a well-defined inference population. However, if the populations change over time, then a sample that is representative of the population in one year may be less representative (compositionally similar) in the following year, which has implications for the relevance of a study's results. One approach that researchers can take to design studies to facilitate generalizations (potentially over time) is to include covariates that change over time in the sampling and recruitment plan. This can be done with sample selection methods based on propensity scores (Tipton et al., 2014), where the propensity scores are used to stratify the inference population and recruitment is conducted within each stratum. Under this approach, researchers can include different types of covariates, including ones that may change quickly over time, in the propensity score model so that the sampling plan incorporates information about variables that can affect the generalizability of the study's results over time. However, it is important to note that this approach for sampling will depend on the

extent to which researchers can anticipate population changes. Additionally, with respect to generalizability, researchers should prioritize the covariates that are important moderators of treatment effects, in both post-hoc generalization analyses and in the design of studies that facilitate generalizations.

Conclusion

This study explored the implications of compositional changes in a population on the generalizability of a study's findings and on the design and planning of future evaluation studies. Changes in the demographics of a population have implications for research practice since they affect the extent to which a study's results are relevant for students in an inference group. Overall, the findings from SimCalc and Indiana illustrate that the most consistent change was in the racial and demographic composition of students. Although the changes were small in magnitude, schools in both case studies experienced a growth in the proportion of non-white students, which implies that many school settings became more racially diverse. Additionally, the increase in the proportion of ELLs, FRPL students and students at risk suggest that over the course of nearly a decade, school leaders and local policymakers faced a growth in student populations that likely required different resources and support.

In addition to the types of population changes that occurred, both SimCalc and Indiana illustrated that covariates change at different paces and by different magnitudes. For example, in SimCalc, the percentage of Hispanic students increased by 1.81 percentage points in a single year while the proportion of FPRL students increased by 2.5 percentage points. This has implications for generalization since the population changes affect the compositional similarity between the sample and the population. If the assumptions for generalization hold, a treatment effect estimate may generalize to schools in one population, but the estimate may be less precise

and more biased for the same population of schools in the following year. Importantly, although the original sample schools may still be generalizable (compositionally similar) to schools in later populations (such as the final year of the time period), or the generalizability of the study sample may be stronger in later years than earlier years, this does *not* imply that treatment effect estimates would still generalize. This is because the types of covariates that affect variation in treatment effects likely change over time, even for the same inference population, so that the assumptions needed for generalization may not hold across all populations. Although the covariates in the case studies were not moderators of treatment effect heterogeneity, our illustrations of the population changes and with the generalizability statistics provide some insight of the pace at which the generalizability of a study sample can decline. Furthermore, the changes in the observed covariates in our analyses may suggest similar changes in unobserved covariates. If these include variables that satisfy the assumptions for generalization, the patterns of change in the generalizability of a study sample over time may be similar to those seen in our case studies.

Population changes affect aspects of study design. When designing studies to facilitate generalization, it is important to recruit samples that are representative of the target inference population, where representativeness refers to compositional similarity. Population changes affect compositional similarity and consequently, the generalizability of a study's results to individuals in the population. To the extent that researchers can anticipate population changes, information about specific covariates that change over time can be incorporated in sampling approaches to ensure that the sample is still representative of the target population. Importantly, if the covariates satisfy the assumptions of generalization, researchers can include them in the

sampling and recruitment plans to ensure that the study sample is generalizable to the inference population, potentially over time.

While the findings in our study provided insight into population changes, there were several limitations and assumptions. First, our discussion of compositional changes was solely based on the covariates used in the original generalization studies (Chan, 2017; Tipton et al., 2017; O'Muircheartaigh & Hedges, 2014; Tipton, 2013). The trends among the populations and generalizability statistics may potentially differ if an alternative set of covariates were used. However, the variables provided information on a variety of school characteristics that offered a general (though not exhaustive) perspective on population changes over time. While comparisons can be made with additional variables, the covariates included in our empirical examples captured several important changes among the populations of students and teachers. Second, while covariates that satisfy sampling ignorability are the most important for generalization, we found no empirical evidence that the covariates used were moderators of treatment effect heterogeneity. However, our discussion of population changes still raises an important consideration for researchers. While the covariates in our case studies did not noticeably affect treatment effect variation, they may be correlated with the true moderators of the intervention effect and an understanding of how they change may still be useful in understanding how the moderators change. Thus, our analyses of population changes may still be useful for understanding how generalizations of study results potentially change over time. Finally, the findings from our study were based on two empirical examples and as a result, the trends observed in SimCalc and Indiana do not necessarily apply to other populations. However, it is important to note that despite differences in context, the populations in both case studies experienced similar changes in student demographics and this trend was also observed among

the subpopulations. Thus, despite the limited number of case studies, the similarity in the types of population changes suggest possible sources of change in other populations. However, future research should continue to explore additional case studies to assess the types of changes that can take place in other populations. The observations from the case studies will provide researchers and policymakers with important information on the changes that affect school communities and the implications of these changes for future studies.

References

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of Calcutta Mathematical Society, 35,* 99 – 109.

Buchanan, A. L., Hudgens, M. G., Cole, S. R., Mollan, K. R., Sax, P. E., Daar, E. S., ... & Mugavero, M. J. (2018). Generalizing evidence from randomized trials using inverse probability of sampling weights. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *181*(4), 1193.

Chan, W. (2017). Partially identified treatment effects for generalizability. *Journal of Research on Educational Effectiveness*, *10*(3), 646 – 669.

Chan, W. (2018). Applications of small area estimation to generalization with subclassification by propensity scores. *Journal of Educational and Behavioral Statistics, 43*(2), 182 – 224.

Chan, W. (2020). The effect of distributional overlap on the precision gain of bounds under subclassification. Forthcoming in *The American Journal of Evaluation*.

Cole, S. R., & Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: The ACTG 320 trial. *American Journal of Epidemiology*, *172*(1), 107 – 115.

Djebbari, H., & Smith, J. (2008). Heterogeneous impacts of PROGRESA. *Journal of Econometrics, 145,* 64 – 80.

Kaput, J. J. (1997). Rethinking calculus: Learning and thinking. *The American Mathematical Monthly*, *104*(8), 731 – 737.

Kern, H.L., Stuart, E.A., Hill, J., & Green, D.P. (2016). Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness, 9*(1), 103 – 127.

Konstantopoulos, S., Miller, S., & van der Ploeg, A. (2013). The impact of Indiana's system of interim assessments on mathematics and reading achievement. *Educational Evaluation and Policy Analysis*, 35(4), 481 – 499.

O'Muircheartaigh, C., & Hedges, L. V. (2014). Generalizing from unrepresentative experiments: a stratified propensity score approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63, 195 – 210.

Olsen, R. B., Orr, L. L., Bell, S. H., & Stuart, E. A. (2013). External validity in policy evaluations that choose sites purposively. *Journal of Policy Analysis and Management*, *32*(1), 107 – 121.

Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., ... & Gallagher, L. P. (2010). Integration of technology, curriculum, and professional development for

advancing middle school mathematics: Three large-scale studies. *American Educational Research Journal*, *47*(4), 833 – 878.

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

Rubin, D. B. (1977). Assignment to Treatment Group on the Basis of a Covariate. *Journal of Educational Statistics*, *2*(1), 1 – 26.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, *6*(1), 34–58.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, 75(371), 591 – 593.

Rubin, D. B. (1986). Statistics and causal inference: Comment: Which ifs have causal answers. *Journal of the American Statistical Association*, 81, 961 – 962.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Stuart, E.A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*(1), 1 – 21.

Stuart,E.A., Cole,S.R., Bradshaw,C.P., & Leaf, P. J. (2011).The use of propensity scores to assess the generalizability of results from randomized trials. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 174, 369 – 386.

Texas Education Agency (2007). Standards for the Operation of School District Disciplinary Alternative Education Programs. Retrieved from http://ritter.tea.state.tx.us/rules/tac/chapter103/ch103cc.html

Texas Education Agency (2018). Enrollment in Texas public schools, 2017 – 2018. Retrieved from https://tea.texas.gov/sites/default/files/enroll_2017-18.pdf

Tipton, E. (2013). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties, and contexts. *Journal of Educational and Behavioral Statistics*, *38*(3), 239 – 266.

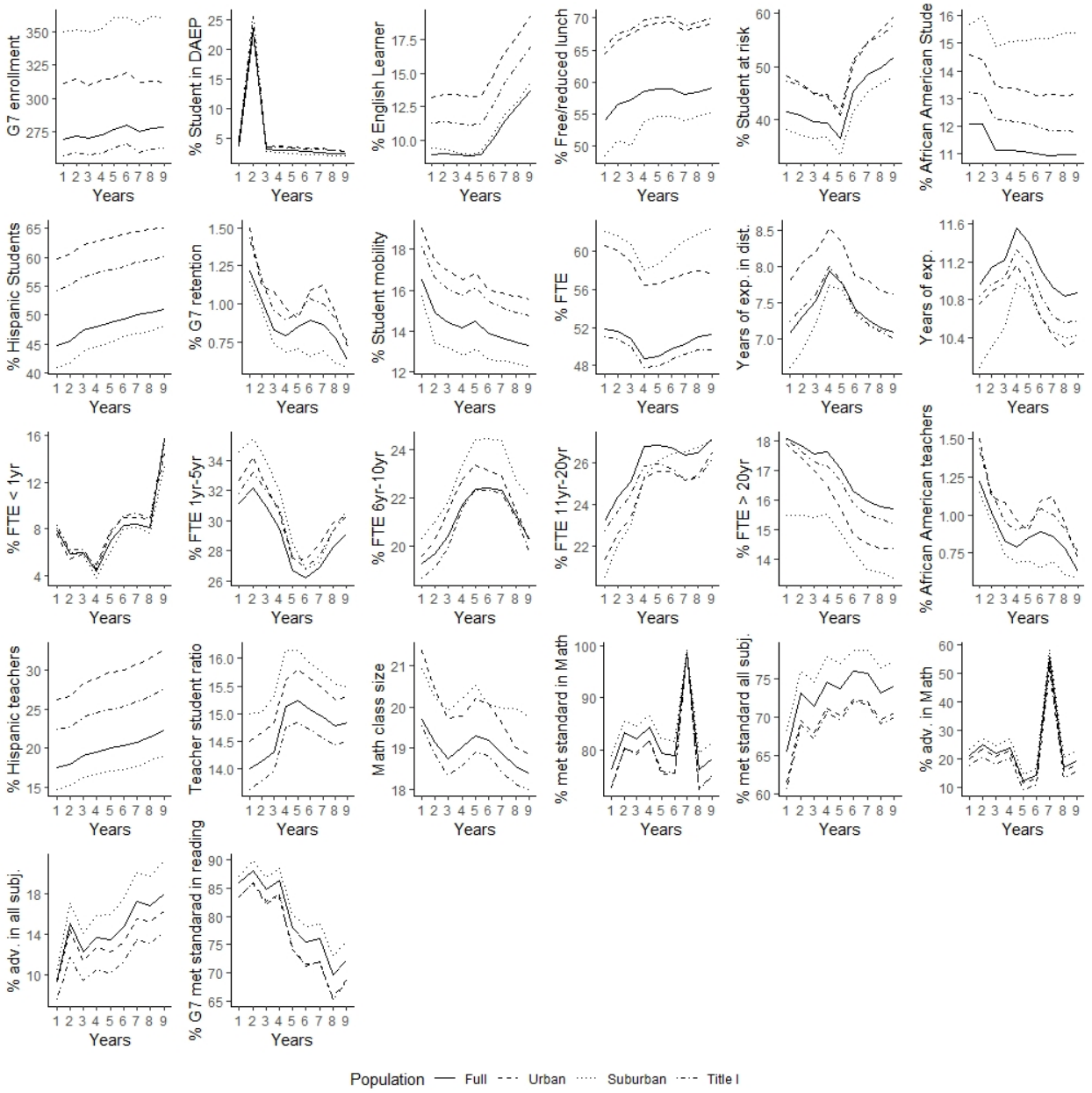
Tipton, E. (2014). How generalizable is your experiment? Comparing a sample and population through a generalizability index. *Journal of Educational and Behavioral Statistics*, 39, 478 – 501.

Tipton, E., Hedges, L.V., Vaden-Kiernan, M., Borman, G., Sullivan, K., & Caverly, S. (2014). Sample selection in randomized experiments: a new method using propensity score stratified sampling. *Journal of Research on Educational Effectiveness, 7*(1), 114 – 135.

Tipton, E., Hallberg, K., Hedges, L. V., & Chan, W. (2017). Implications of small samples for generalization: Adjustments and rules of thumb. *Evaluation Review*, *41*(5), 472–505.

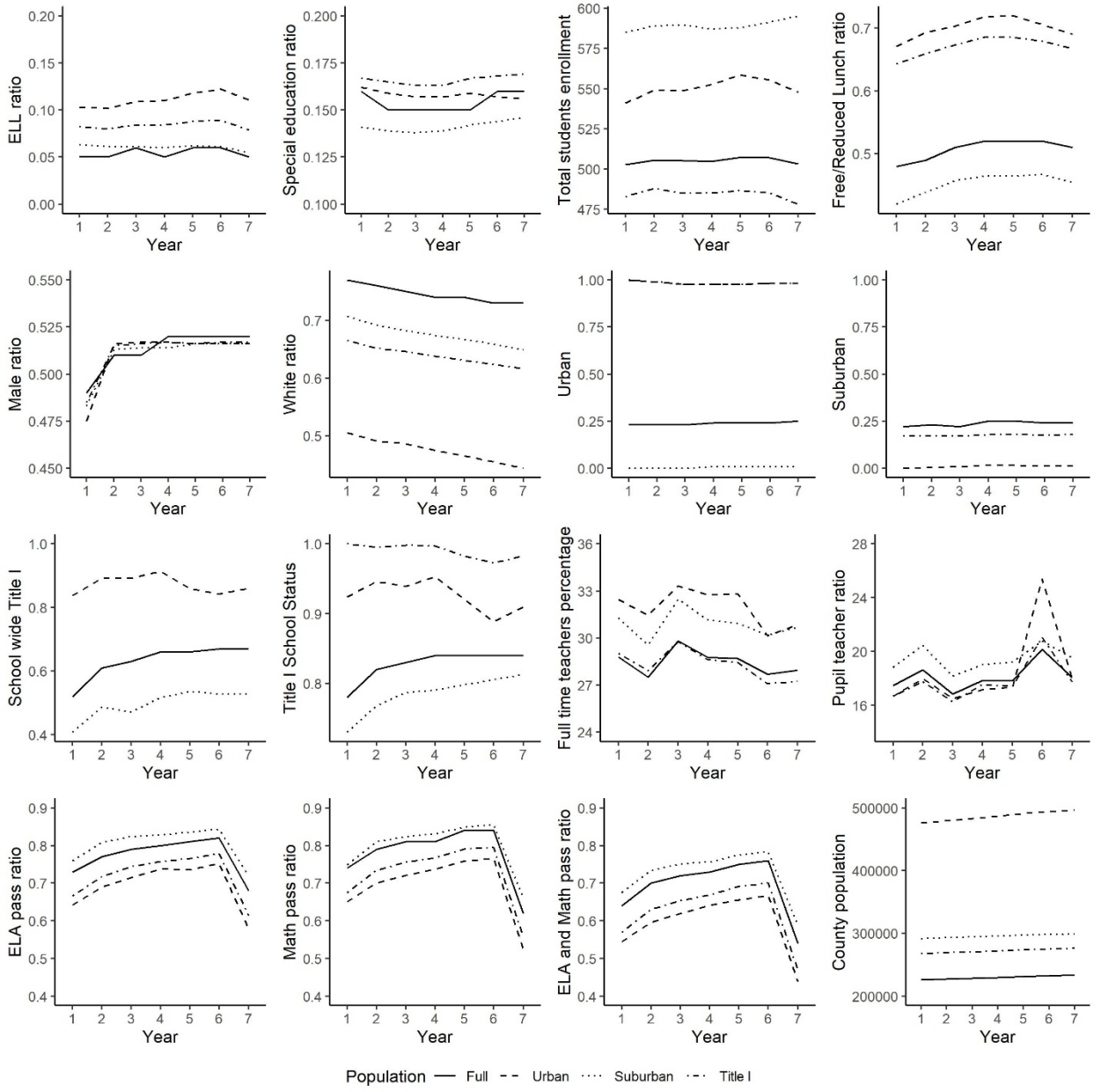
Acknowledgements
The authors thank the associate editor, Dr. Luke Miratrix, and three anonymous reviewers for their helpful feedback and comments, which greatly improved the original manuscript. The authors also thank SRI International for the SimCalc data.

Figure 1. Changes in Covariates for SimCalc



Note: Year 1 refers to 2008 – 09, Year 2 refers to 2009 – 2010, Year 3 refers to 2010 – 2011, Year 4 refers to 2011 – 2012, Year 5 refers to 2012 – 2013, Year 6 refers to 2013 – 2014, Year 7 refers to 2014 – 2015, Year 8 refers to 2015 – 2016, and Year 9 refers to 2016 – 2017.

Figure 2. Changes in Covariates for Indiana



Note: Year 1 refers to 2009 – 2010, Year 2 refers to 2010 – 2011, Year 3 refers to 2011 – 2012, Year 4 refers to 2012 – 2013, Year 5 refers to 2013 – 2014, Year 6 refers to 2014 – 2015, and Year 7 refers to 2015 – 2016.

Figure 3. Generalizability Index for SimCalc

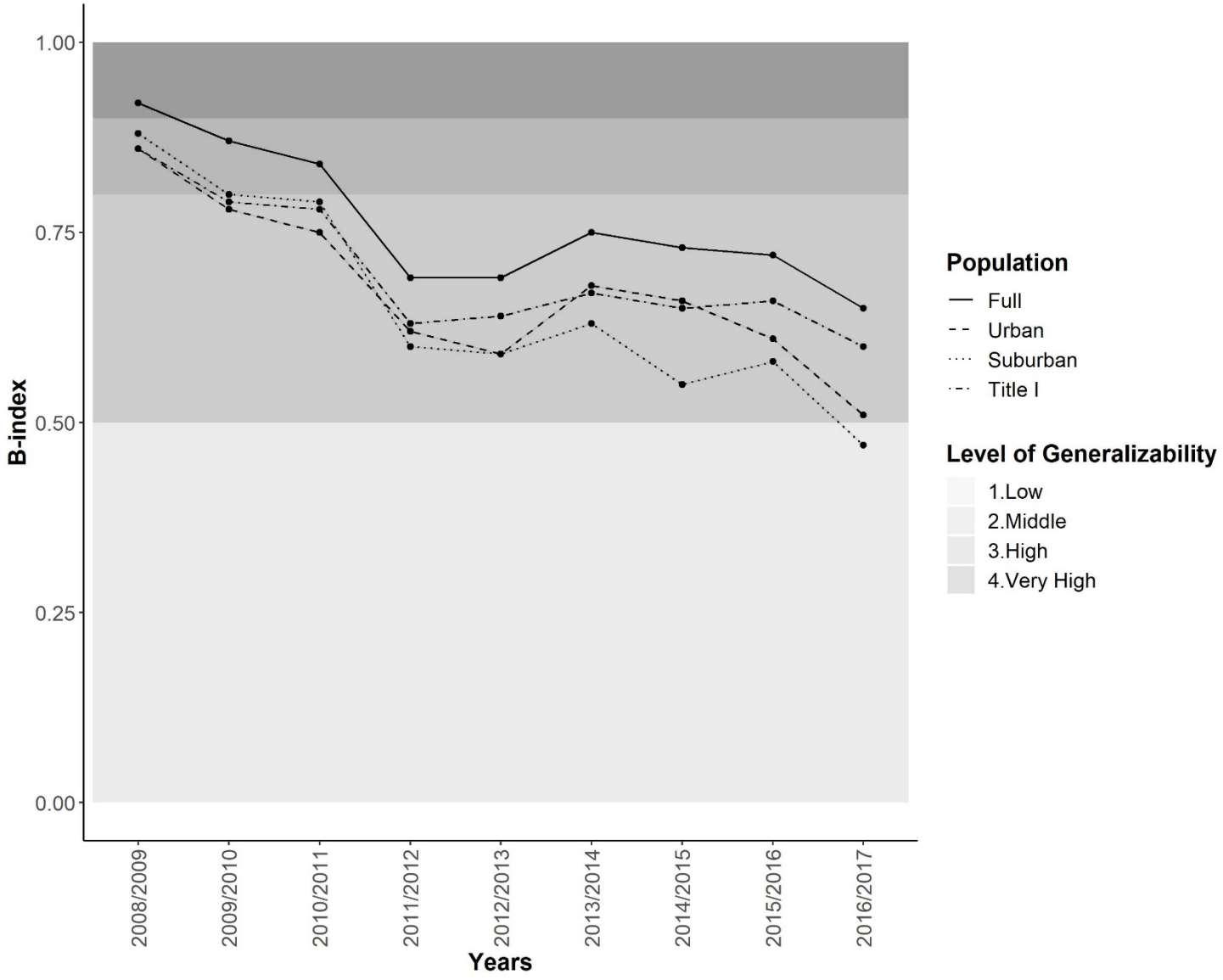


Figure 4. Overlap for SimCalc

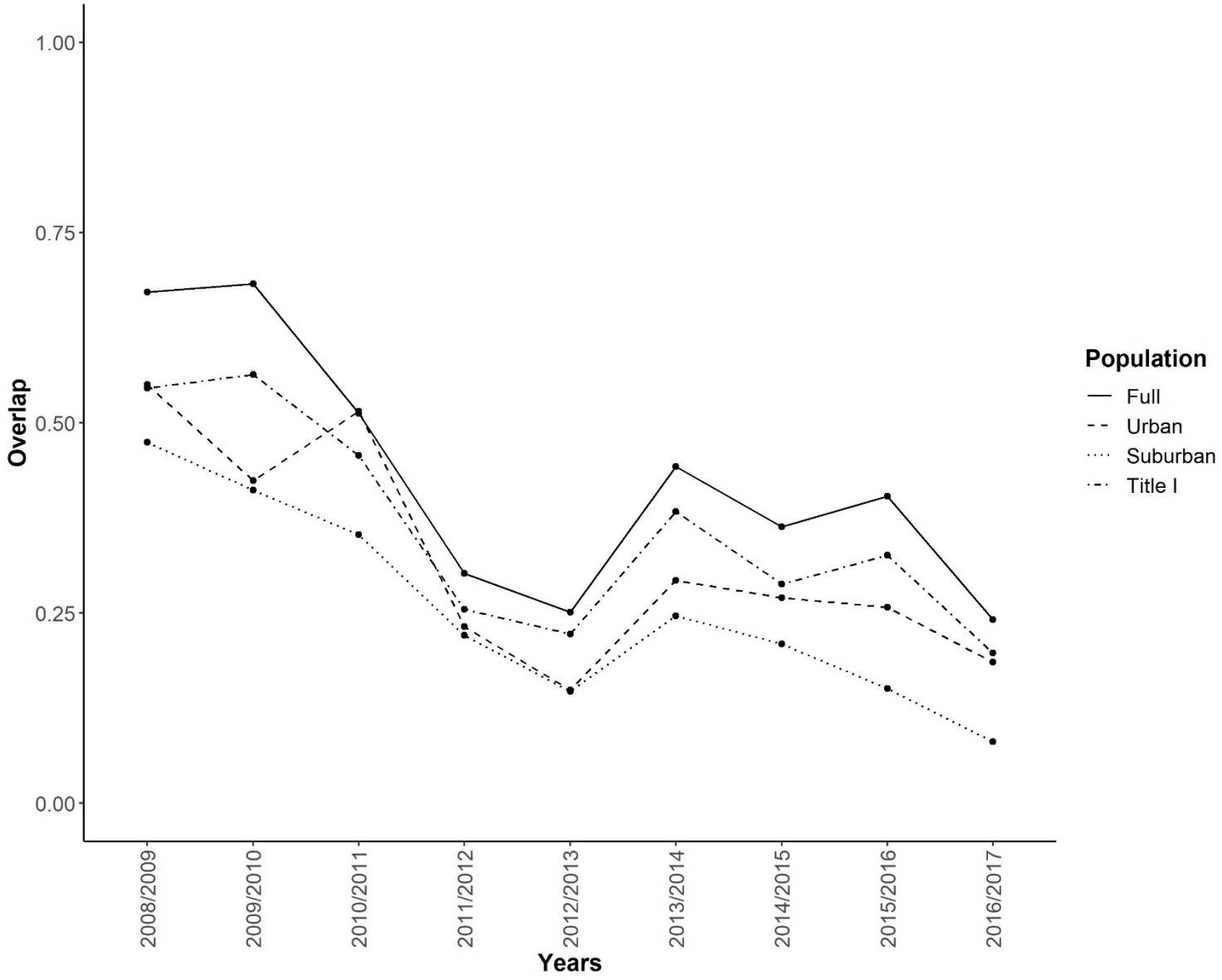


Figure 5. Generalizability Index for Indiana

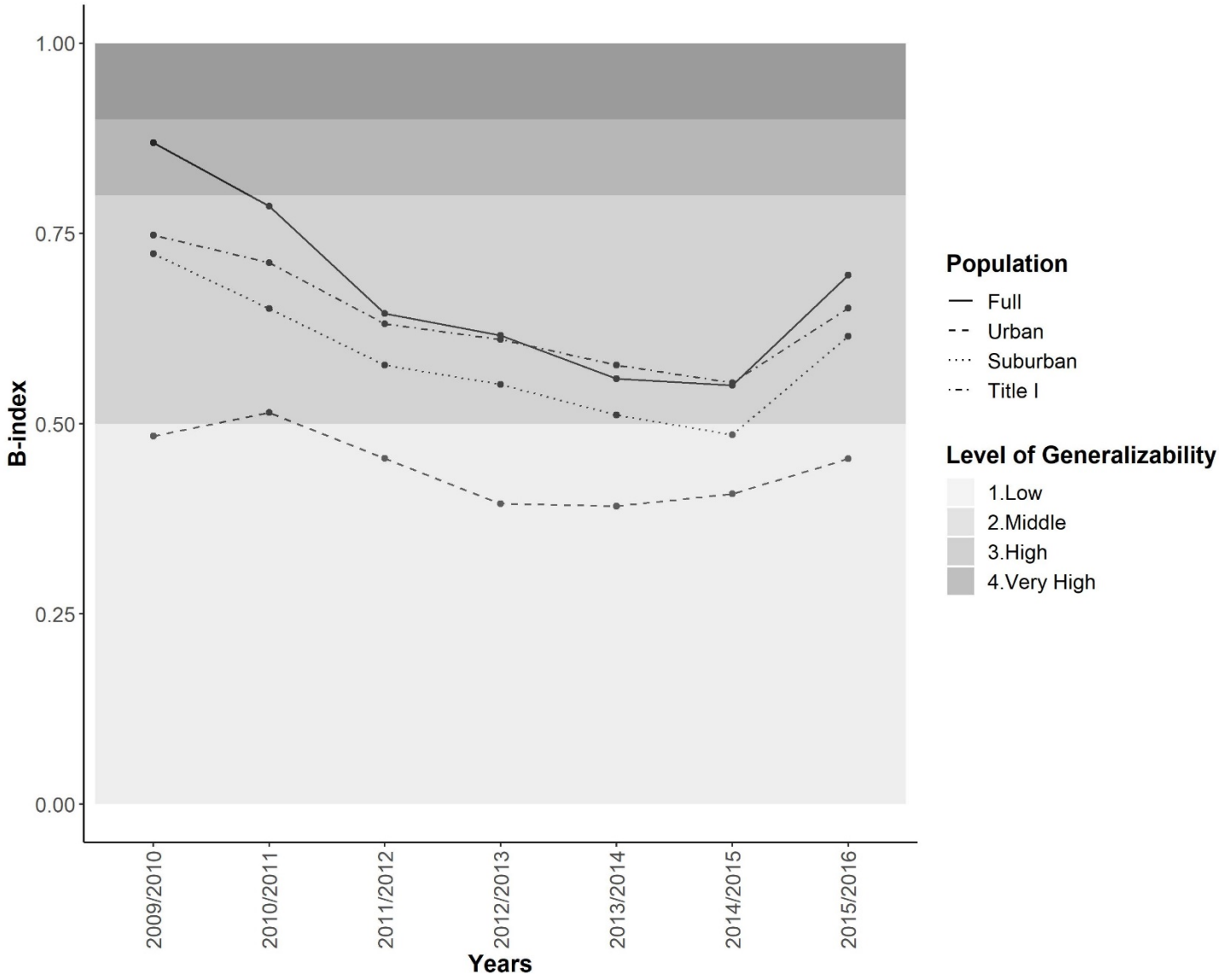


Figure 6. Overlap for Indiana

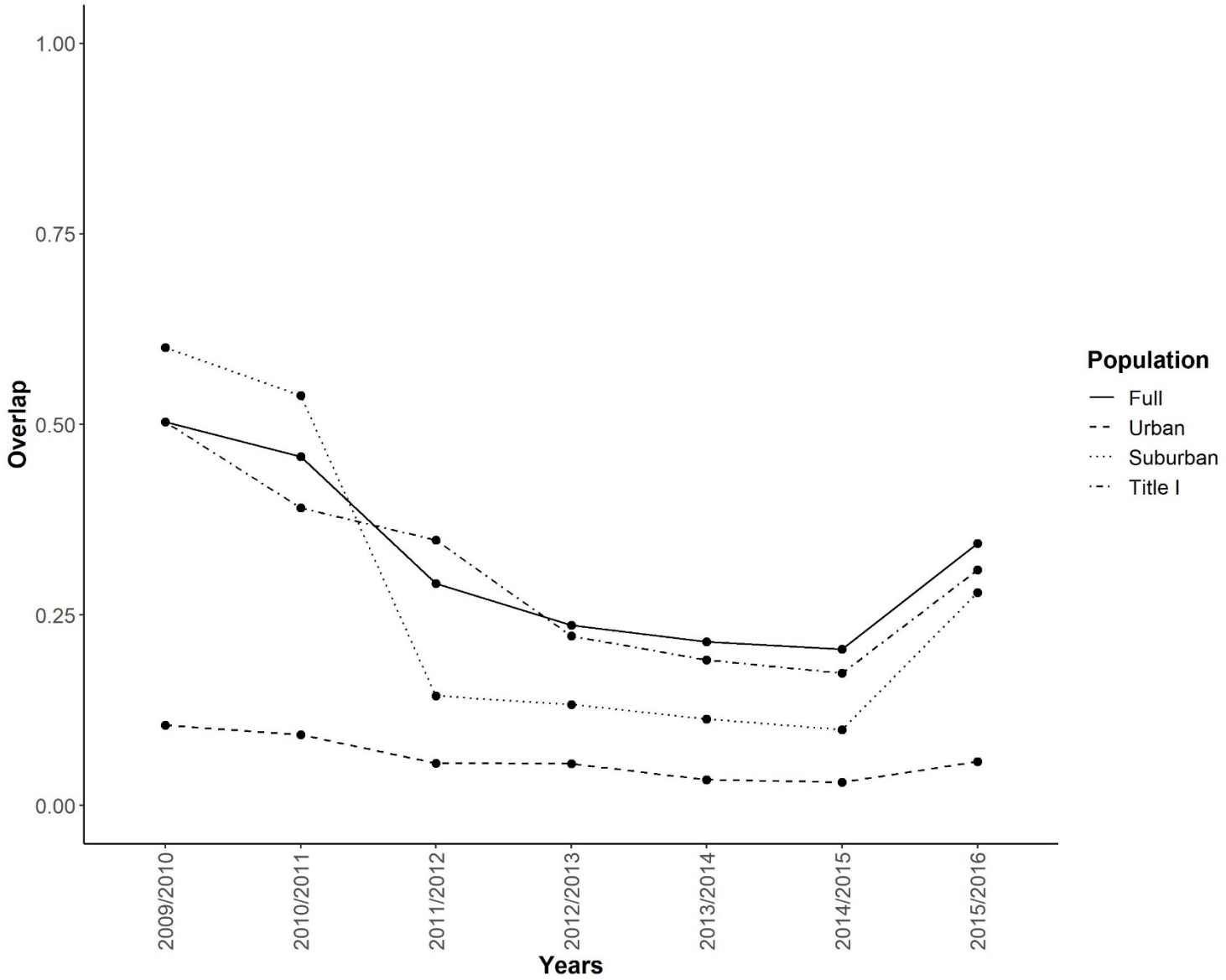


Table 1. Covariate means for SimCalc and Texas population schools from 2008-09 to 2016-17

	Sim-Calc (*n = 63)*	2008-2009 (*N = 936*)	2009-2010 (*N = 936*)	2010-2011 (*N = 936*)	2011-2012 (*N = 936*)	2012-2013 (*N = 936*)	2013-2014 (*N = 936*)	2014-2015 (*N = 936*)	2015-2016 (*N = 936*)	2016-2017 (*N = 936*)
7th Grader enrollment (mean)	296.86	269.05	271.56	269.61	272.30	276.70	279.49	274.86	277.39	278.54
Students who are in disciplinary alternative education program (DAEP) (%)	4.09	3.49	22.83	2.98	2.89	2.80	2.61	2.55	2.33	2.31
English Learners (%)	11.22	8.92	8.98	8.95	8.83	8.95	10.08	11.47	12.51	13.65
Free/Reduced Lunch (%)	54.38	53.96	56.49	57.20	58.61	58.94	59.01	58.07	58.48	59.12
Students at risk (%)	41.34	41.57	40.89	39.63	39.48	36.56	45.34	48.57	49.71	51.80
African American students (%)	8.52	12.08	12.06	11.11	11.11	11.07	10.98	10.91	10.96	10.96
Hispanic students (%)	49.74	44.69	45.55	47.36	48.00	48.66	49.34	49.97	50.44	51.08
Grade 7 retention (%)	0.97	1.22	1.02	0.83	0.79	0.85	0.89	0.86	0.78	0.63
Student mobility	15.60	16.50	14.93	14.44	14.18	14.50	13.87	13.68	13.48	13.28
Full Time Equivalent (FTE) teachers (mean)	56.08	51.81	51.58	50.79	48.70	48.91	49.75	50.27	50.99	51.26
Years of teaching within district (mean)	7.06	7.08	7.32	7.53	7.94	7.76	7.42	7.26	7.15	7.10

Years of teaching experience (mean)	10.50	10.95	11.14	11.22	11.56	11.41	11.11	10.93	10.84	10.87
FTE Beginning Teachers (%)	9.50	7.83	5.94	5.98	4.36	7.01	8.33	8.46	8.18	15.70
FTE Teachers 1-5 years (%)	31.30	31.14	32.15	30.93	29.50	26.72	26.20	26.91	28.25	29.11
FTE Teachers 6-10 years (%)	20.67	19.26	19.67	20.39	21.73	22.33	22.41	22.29	21.27	20.24
FTE Teachers 11-20 years (%)	22.30	23.21	24.37	25.14	26.79	26.85	26.78	26.38	26.52	27.18
FTE teachers 20 + years (%)	16.71	18.07	17.87	17.57	17.62	17.09	16.29	15.95	15.78	15.70
African American teachers (%)	5.42	1.22	1.02	0.830	0.79	0.85	0.89	0.86	0.78	0.63
Hispanic teachers (%)	23.36	17.58	17.88	19.01	19.53	20.01	20.31	20.78	21.46	22.30
Teacher/student ratio (mean)	14.20	14.00	14.15	14.3	15.14	15.24	15.08	14.93	14.78	14.84
Math class size (mean)	20.42	19.70	19.15	18.74	19.04	19.31	19.19	18.88	18.55	18.4
G3-G11 Met standard rate in math (%)	75.60	76.17	83.36	82.06	84.29	79.29	78.96	98.83	76.05	78.27
G3-G11 met standard rate in all tests (%)	64.57	65.38	73.10	71.45	74.64	73.75	75.95	75.76	73.21	73.95

G3-G11 advanced performance in math (%)	21.57	21.06	24.81	21.82	23.93	12.24	14.02	55.60	16.93	19.10
G3-G11 advanced performance in all tests (%)	9.62	9.45	15.04	12.26	13.65	13.47	14.79	17.31	16.78	18.07
G7 met standard rate reading (%)	85.79	86.01	88.14	84.79	86.32	78.05	75.53	76.09	69.65	72.24
Students who are proficient in 7th grade mathematics (%)	75.61	72.80	-	-	-	-	-	-	-	-
% of students with commended performance, grades 3 – 11, reading	8.71	8.71	-	-	-	-	-	-	-	-

Note: SimCalc took place in the 2008 – 2009 academic year. The populations defined in each academic year comprise all public schools in Texas that serve 7th grade students. The last two rows of the table correspond to the two covariates that were in the original generalization studies but were not part of the analyses of this study. The covariates in bold in the table are new variables added to the analysis.

Table 2. Covariate means for Indiana CRT and population from 2009 – 2010 to 2015 – 2016

Covariate Descriptions	CRT (*n = 41*)	2009-2010 (*N = 1225*)	2010-2011 (*N = 1225*)	2011-2012 (*N = 1225*)	2012-2013 (*N = 1225*)	2013-2014 (*N = 1225*)	2014-2015 (*N = 1225*)	2015-2016 (*N = 1225*)
ELL ratio	0.03	0.05	0.05	0.06	0.05	0.06	0.06	0.05
Special education ratio	0.17	0.16	0.15	0.15	0.15	0.15	0.16	0.16
Total student enrollment	443.12	502.77	505.58	505.22	504.90	507.46	507.21	503.33
Free/Reduced Lunch ratio	0.46	0.48	0.49	0.51	0.52	0.52	0.52	0.51
Male ratio	0.50	0.49	0.51	0.51	0.52	0.52	0.52	0.52
White ratio	0.89	0.77	0.76	0.75	0.74	0.74	0.73	0.73
Urban	0.07	0.23	0.23	0.23	0.24	0.24	0.24	0.25
Suburb	0.05	0.22	0.23	0.22	0.25	0.25	0.24	0.24
Schoolwide Title I	0.54	0.52	0.61	0.63	0.66	0.66	0.67	0.67
Title I School Status	0.83	0.78	0.82	0.83	0.84	0.84	0.84	0.84
Full time teachers percentage	25.14	28.8	27.51	29.81	28.77	28.69	27.70	27.94
Pupil teacher ratio	17.86	17.46	18.64	16.85	17.84	17.81	20.16	18.05
ELA pass ratio	0.74	0.73	0.77	0.79	0.80	0.81	0.82	0.68
Math pass ratio	0.76	0.74	0.79	0.81	0.81	0.84	0.84	0.62
ELA and Math pass ratio	0.66	0.64	0.70	0.72	0.73	0.75	0.76	0.54
County population	91318.00	225825.79	227175.11	228353.68	229738.45	231605.12	232662.13	233464.11
2008 – 2009 Attendance	96.44	96.38	-	-	-	-	-	-
2008 – 2009 ELA Test Scores	19.41	18.82	-	-	-	-	-	-
2008 – 2009 Math Test Scores	16.48	16.46	-	-	-	-	-	-

ELL refers to English Language Learner. The Indiana CRT took place in the 2009 – 2010 academic year. The populations defined in each academic year comprise all public K – 8 schools in Indiana with some exclusions. The last three rows of the table correspond to the covariates that were in the original generalization studies but were not part of the analyses of this study. The variables in bold are new covariates that were not part of the original study but were added to the analysis of this study.