**Title:** How generalizable are study samples over time?

**Authors:**

Wendy Chan

Graduate School of Education, University of Pennsylvania

Email: wechan@upenn.edu

Jimin Oh

Graduate School of Education, University of Pennsylvania

Email: ohjimin@upenn.edu

Peihao Luo

Graduate School of Education, University of Pennsylvania

Email: peihao@upenn.edu

**Background/Context:**

Experimental studies are considered the gold standard for evaluating the causal impact of a program or intervention. For the past 8 years, statisticians have developed methods to formally address a growing interest in the *generalizability* of experimental results; namely, the expected intervention impact for a specified population of inference (Stuart et al., 2011; Tipton, 2013; O'Muircheartaigh & Hedges, 2014; Chan, 2017). These methods focus on improving the generalizability of results when the study sample is not randomly selected from the population. These methods are based on propensity scores, which estimate the conditional probability of study selection given a set of observable covariates. Since their development, propensity scores have made an important contribution to improving generalizations through their use in matching and reweighting approaches.

**Purpose/Research Questions:**

Generalization research thus far has focused on populations that are defined in a cross-sectional setting; namely, for educational studies, the target population is typically defined for the same academic year. The purpose of this project is to evaluate whether the generalizability of a study's results potentially changes over time. Specifically, we consider the context in which there are multiple populations of inference, each corresponding to a different academic year, and assess the compositional similarity (or difference) between the study sample and each population. Our project addresses the following research questions:

1. How does the composition of a population change, in relation to a study sample, over time and what are the implications for the generalizability of the study's results?
2. What is the "pace" at which the population composition changes?

**Setting/Population:**

Our project uses data from SimCalc, a cluster-randomized trial (CRT) that evaluated the effectiveness of a technology-based curriculum on mathematics achievement among seventh-grade students in Texas (Roschelle et al., 2010). The population of inference is specified as all public Pre-K to Grade 12 schools in Texas. The original study sample consisted of 92 schools, of which 45 were randomized to treatment (SimCalc) and the other 47 to control.

**Data Collection and Analysis:**

We constructed a population data frame using the Academic Excellence Indicator System and Texas Academic Performance Report. Data was collected on 26 covariates, which included aggregate measures of demographic and academic achievement variables. To address our research questions, the population data frame ranged from 2008 – 2009 (the year in which SimCalc was conducted) to 2016 – 2017 where each academic year represents a separate inference population. Because of initial inconsistencies in the data, the final dataset (across all years) comprised of 936 population schools and 63 study schools.

We used the generalizability index (*B*-index; Tipton, 2014) and the distributional overlap in the estimated propensity scores to assess the similarity between the study samples and each population of inference. The *B*-index ranges from 0 to 1 and following Tipton (2014), we used the following four ranges to assess generalizability: $1.00 \leq B \leq .90$; very high; $.80 \leq B \leq .90$; high; $.70 \leq B \leq .50$; middle; $B < .50$; low. Overlap refers to the proportion of population schools

whose propensity scores lie in the range of the propensity scores of the sample. For all academic years, we estimated the propensity scores using logistic regression.

The analysis was conducted in three stages. First, generalizability statistics (the *B*-index and overlap) were estimated for all nine years using a propensity score model based on the original 26 covariates. Because of changes in the Texas state exam in 2011 – 2012, there was a significant amount of variability in the propensity scores for certain years. As a result, our second stage involved refitting the propensity score model on a subset of covariates and re-computing the generalizability statistics. Lastly, the third stage considered the generalizability of the study sample to a smaller population that consisted of urban schools in Texas for each academic year. The purpose of this last comparison is to assess whether the values and changes in the generalizability statistics differ at a more local level of generalization.

**Preliminary Findings:**

Tables 1 – 3 provide the results for each analysis. From Tables 1 and 2, the generalizability between the SimCalc sample (63 schools) and population is highest in 2008 – 2009, the same year in which the study was conducted. Table 1 illustrates that there are convergence issues in the propensity score model, likely due to changes in the Texas statewide testing system. Variables such as the percentage of students in the disciplinary alternative education program (DAEP) have different distributions in specific years, where the median value is 0.20 for 2009 – 2010, but it is 0.03 for the other years (excluding 2009 – 2010). Table 2 provides the results when these variables are excluded. The trends in the *B*-index and overlap suggest that the generalizability of the SimCalc sample immediately declines in the years following the study, but the largest decline begins to happen three years after the study. In particular, the *B*-index decreases from 0.92 to 0.69 in 2011 – 2012. Interestingly, the *B*-index remains in the "Middle" range even after eight years as seen in 2016 – 2017. These changes are also reflected in the overlap whose value is the highest in 2008 – 2009 at 0.92 and drops to the 0.70 range in 2016 – 2017.

The inference populations for Table 3 consisted of all Texas urban schools. Like Table 2, the *B*-index and overlap between the SimCalc sample and the urban population are highest in 2008 – 2009. However, the generalizability statistics are lower in this population compared to the populations in Table 2. Additionally, the values of the generalizability statistics decrease at a faster rate for the populations in Table 3, implying that there is a growing dissimilarity in propensity scores between the SimCalc schools and Texas urban schools with each successive year. Furthermore, the overlap in propensity score distributions is nearly 58% smaller in 2016 – 2017 for the urban schools compared to its value for the same year in Table 2. However, despite the faster changes, the *B*-index is still considered "Middle," albeit on the lower end of the range, after eight years in 2016 – 2017.

Table 1. Generalizability statistics (26 covariates)

| Year | SimCalc 0809 | SimCalc 0910 [ab] | SimCalc 1011 | SimCalc 1112 | SimCalc 1213 [ab] | SimCalc 1314 [ab] | SimCalc 1415 [ab] | SimCalc 1516 [ab] | SimCal 1617 [ab] |
|---|---|---|---|---|---|---|---|---|---|
| *B*-index | 0.91 | 0.34 | 0.52 | 0.41 | 0.00 | 0.00 | 0.17 | 0.00 | 0.00 |
| Decision | Very High | Low | Middle | Low | Low | Low | Low | Low | Low |
| Overlap | 0.9119 | 0 | 0.5843 | 0.7850 | 0 | 0 | 0 | 0 | 0 |

a. Fitted probabilities numerically 0 or 1 occurred

b. The algorithm did not converge.

Note. The propensity score model is based on the following covariates: percentage of English learners, percentage of African American Students or Hispanic students, percentage of students who are at risk, percentage of students who are in disciplinary alternative education program(DAEP), total number of full time equivalent teacher (FTE), percentage of beginning/1-5 years/ 6-10 years/ 11-20 years/ more than 20 years teachers, average teachers' years of experience within school and as total, percentage of African American teacher or Hispanic teacher, school size, grade 7 retention rate, student mobility rate, teacher/student ratio, average size of math class, percentage of students who meet the standard in all subjects/mathematics, percentage of students who meet the advanced standard in all subjects/mathematics, percentage of grade 7 students who meet the standard in reading. "Decision" refers to the category for the range in which the *B*-index falls.

Table 2. Generalizability statistics of covariates balance (20 covariates)

| Year | SimCalc 0809 | SimCalc 0910 | SimCalc 1011 | SimCalc 1112 | SimCalc 1213 | SimCalc 1314 | SimCalc 1415 | SimCalc 1516 | SimCal 1617 |
|---|---|---|---|---|---|---|---|---|---|
| *B*-index | 0.92 | 0.87 | 0.84 | 0.69 | 0.69 | 0.75 | 0.73 | 0.72 | 0.65 |
| Decision | Very High | High | High | Middle | Middle | Middle | Middle | Middle | Middle |
| Overlap | 0.9119 | 0.8789 | 0.8749 | 0.7758 | 0.7838 | 0.8078 | 0.7708 | 0.7958 | 0.7578 |

Note. The following covariates were excluded from this model: percentage of students who are in disciplinary alternative education program (DAEP), percentage of students who meet the standard in all subjects/mathematics, percentage of students who meet the advanced standard in all subjects/mathematics, percentage of grade 7 students who meet the standard in reading.

Table 3. Generalizability statistics for the population of urban schools (20 covariates)

| Year | SimCalc 0809 | SimCalc 0910 | SimCalc 1011 | SimCalc 1112 | SimCalc 1213 | SimCalc 1314 | SimCalc 1415 | SimCalc 1516 | SimCal 1617 |
|---|---|---|---|---|---|---|---|---|---|
| *B*-index | 0.86 | 0.78 | 0.75 | 0.62 | 0.59 | 0.68 | 0.66 | 0.61 | 0.51 |
| Decision | High | Middle | Middle | Middle | Middle | Middle | Middle | Middle | Middle |
| Overlap | 0.8131 | 0.7828 | 0.7626 | 0.5758 | 0.5227 | 0.5758 | 0.4697 | 0.3838 | 0.3207 |

Note. The populations of inference consist of 396 urban schools located in the major cities of Texas.

## References


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

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*(2), 195 – 210.

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.

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*(2), 369 – 386.

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? An index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics*, *39*(6), 478 – 501.