How Generalizable are Study Samples Over Time?

Wendy Chan, Jimin Oh, Peihao Luo

Human Development and Quantitative Methods Division, Graduate School of Education, University of Pennsylvania

Background

- Experimental studies are considered the gold standard for evaluating the causal impact of a program or intervention.
- Statisticians have developed methods to formally address a growing interest in the *generalizability* of experimental results; namely, the extent to which the results from the studies are applicable to individuals in a specified population of inference (Stuart et al., 2011; Tipton, 2013; O'Muircheartaigh & Hedges, 2014; Chan, 2017).
- Generalizability research thus far has focused on populations which have typically been specified in a cross-sectional context; namely, the populations are defined in the same year/academic year in which the study took place or there is a single population defined at a fixed time point.
- To date, little research has been done to examine how the generalizability of a study's results changes over time when considering a single population of inference over multiple time points.

Research Questions

1. To what extent does the generalizability of a study sample change when the population of inference is defined at different time points?
2. What is the "pace" at which the generalizability of a study sample changes?
3. Does the pace of change differ based on the definition of the population? In other words, is the pace the same for subpopulations within the broader population of interest?

Setting and 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

- Data was obtained from Academic Excellence Indicator System (AEIS) and Texas Academic Performance Report (TAPR)
  - Covariates include aggregate measures of student and teacher demographics, school features, and academic achievement
- However, because of initial inconsistencies in the data, the final dataset (across all years) comprised of 936 population schools and 63 study sample schools.
- Inference population: All eligible schools in Texas starting from 2008 – 2009 (the year in which SimCalc was conducted) to 2016 – 2017. Each academic year represents a separate inference population.

Methodology

- 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 sample and each population of inference.
  - The *B*-index : *B*-index uses the density functions of estimated propensity score logits to quantify the similarity in propensity score distributions between the sample and the population.
  - Overlap : the proportion of population schools whose propensity scores lie in the range of the propensity scores of the sample.
- For each school, let $Z = 1$ if a school was in the sample and let $\boldsymbol{X}$ represent a vector of observable covariates. The sampling propensity score is given by $s(\boldsymbol{X}) = \Pr(Z = 1|\boldsymbol{X})$.
- Propensity scores were estimated using logistic regression.

Analysis

The analysis was conducted in three stages:

1. We estimated the *B*-index and distributional overlap between the SimCalc sample and each inference population for all nine years using a propensity score model based on the original 26 covariates.
2. We refit the propensity score model on a subset of covariates and re-estimated the *B*-index and overlap.
3. We estimated the generalizability statistics on several subpopulations that consisted of urban, suburban, and School-wide Title I schools in Texas for each academic year.

Results

Generalizability statistics (26 covariates)

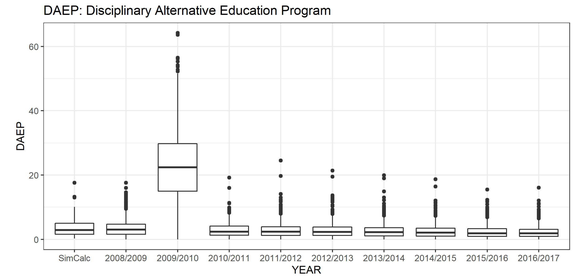
	2008-2009	2009-2010 [ab]	2010-2011	2011-2012	2012-2013 [ab]	2013-2014 [ab]	2014-2015 [ab]	2015-2016 [ab]	2016-2017 [ab]
B-index	0.91	0.34	0.52	0.41	0.00	0.00	0.17	0.00	0.00
Generalizability	Very High	Low	Middle	Low	Low	Low	Low	Low	Low
Overlap	91.19	0	58.43	78.50	0	0	0	0	0

a. Fitted probabilities numerically 0 or 1 occurred
b. The algorithm did not converge

Note: "Generalizability" refers to the category for the range in which the *B*-index falls (Tipton, 2014).

Challenges with particular covariates

- The distribution for the DAEP variable was different in one year, which led to complications in the propensity score model.



- Academic achievement measure change: Changes in the Texas state exam in 2012 – 2013, from TAKS to STAAR.
- Additional analyses were done before and after the state exam change.

Generalizability statistics (20 covariates)

Full population

	2008-2009	2009-2010	2010-2011	2011-2012	2012-2013	2013-2014	2014-2015	2015-2016	2016-2017
B-index	0.92	0.87	0.84	0.69	0.69	0.75	0.73	0.72	0.65
Generalizability	Very High	High	High	Middle	Middle	Middle	Middle	Middle	Middle
Overlap	91.19	87.89	87.49	77.58	78.38	80.78	77.08	79.58	75.78

Subpopulation : Urban

	2008-2009	2009-2010	2010-2011	2011-2012	2012-2013	2013-2014	2014-2015	2015-2016	2016-2017
B-index	0.86	0.78	0.75	0.62	0.59	0.68	0.66	0.61	0.51
Generalizability	High	Middle	Middle	Middle	Middle	Middle	Middle	Middle	Middle
Overlap	81.31	78.28	76.26	57.58	52.27	57.58	46.97	38.38	32.07

Note. The populations of inference consist of 333 urban schools located in the major cities of Texas.

Results (Cont'd)

Subpopulation : Suburban

	2008-2009	2009-2010	2010-2011	2011-2012	2012-2013	2013-2014	2014-2015	2015-2016	2016-2017
B-index	0.88	0.80	0.79	0.60	0.59	0.63	0.55	0.58	0.47
Generalizability	High	Middle	Middle	Middle	Middle	Middle	Middle	Middle	Low
Overlap	69.85	69.49	62.50	39.71	44.12	34.56	32.35	33.09	27.57

Note. The populations of inference consist of 209 suburban schools located outside the major cities of Texas.

Subpopulation : School-wide Title I Schools

	2008-2009	2009-2010	2010-2011	2011-2012	2012-2013	2013-2014	2014-2015	2015-2016	2016-2017
B-index	0.86	0.79	0.78	0.63	0.64	0.67	0.65	0.66	0.60
Generalizability	High	Middle	Middle	Middle	Middle	Middle	Middle	Middle	Middle
Overlap	88.68	84.86	83.90	64.12	67.30	55.94	58.25	63.98	59.07

Note. The populations of inference consist of 670 School-wide Title I schools in Texas. School-wide Title I status is given to schools whose populations consist of at least 40% low-income students.

Findings

- 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.
- *B*-index remains in the "middle" range even after eight years as seen in 2016 – 2017. Values of the overlap are consistent with these changes.
- The patterns of changes in the *B*-index and overlap between SimCalc and the subpopulations were similar to the full population where the generalizability was highest in 2008-2009 before a substantial decline takes place at the three-year mark.
- The decline in generalizability happens faster for subpopulations, particularly for the population of suburban schools whose *B*- index values drop to the "low" range in the study.

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.