Building generalizations: Tools for increasing the relevance of your results

Elizabeth Tipton
Teachers College, Columbia University

Presented at IES PI Meeting
December 10, 2015

Relevance

Today I'm going to talk about methods for improving the **relevance** of large-scale experiments.

Questions of **relevance** are inherently questions about **generalizability**.

Improved generalizability = improved relevance.

Overview

Two approaches to increasing relevance:

- Design – build a better sample

- Assessment – Help others understand where the results may apply and where they may not

Overview

Two approaches to increasing relevance:

- Design – build a better sample
  - Tool 1: The Generalizer (planning)
- Assessment – Help others understand where the results may apply and where they may not
  - Tool 2: The Generalizer (assessment)

Overview

Two approaches to increasing relevance:

- Design – build a better sample
  - Tool 1: The Generalizer (planning)
- Assessment – Help others understand where the results may apply and where they may not
  - Tool 2: The Generalizer (assessment)

(Not discussed: Post-hoc estimators)

Background

THE GENERALIZATION PROBLEM

How do we know if results generalize?

We all have informal ideas about generalization.

For example, we might think that results generalize if:

- The sample was selected randomly; or
- The treatment effect is constant (no interactions); or
- The contexts are similar.

Representative samples

Kruskal and Mosteller (1979) wrote a great series of papers on how the term *representative sample* is used.

They found 9 uses in the statistical literature.

The three most relevant here are:

1. **Typical** or ideal cases
2. **Coverage** of the population
3. **Miniature** of the population

Typical cases

One type of representative sample is one composed of typical cases. These are units closest to the modal type.

This kind of sample is homogenous on many covariates. It includes only “average” units.

Informal concept of generalization:

> *Results generalize if they are found to work in the typical unit.*

Coverage

Another type of representative sample is one composed of each type of unit in the population, regardless of frequency.

For example, a study may aim to include at least one African-American, White, Asian, Hispanic, and Native American student.

Informal concept of generalization:

> *Results generalize if they are still found even with such large diversity in the sample.*

Miniature

The third type is a sample that is a miniature of the population. Like coverage, this involves including the same diversity found in the population.

In addition, however, the composition of units should be the same in both the sample and population. That is, the **frequency** of each type is the same.

Informal concept of generalization:

> *Results generalize to a population if they are found to work in sample that is "like" it.*

Miniature

Today I'm going to focus on this third type for an important reason:

> Experiments estimate average treatment effects.

This begs the question, average treatment effect *for whom*?

In order for research to be relevant, we need this to be a policy relevant population.

Probability sampling

When the goal is for the sample to be a miniature of the population, the most obvious approach is to use **probability sampling**.

Probability sampling is great when you can do it.

However,

- It is very rarely done.
- Often there is high non-response.
- Even in big congressionally mandated studies, you can get funny probability samples.

Generalization

In practice, samples are rarely selected from well-defined populations. Instead, generalizations are made **post-hoc**.

For example, a policy maker visits the *What Works Clearinghouse* or reads an article and must decide whether the results of a study are relevant to their policy relevant inference population.

They want to know: Will this work in my state? For Title I schools? For schools like mine?

The problem

Policy makers have **few tools** and **limited information** to make generalizations well.

They likely *don't know* the composition of the study population (and maybe the inference population too) if the study is not explicit.

They *don't know* how to choose what compositional variables might matter.

At best, a policy maker can judge coverage, but not frequency or joint frequency.

The role of the evaluator

Unlike policy makers, those conducting large-scale evaluations:

- Have a rich understanding of the treatment;
- Have a sense of the conditions under which it may work best;
- Understand features of the study;
- Have the relevant data to make comparisons.

For these reasons, I argue that researchers should lead the conversation about generalization.

Tool #1

PLANNING/ DESIGNING FOR GENERALIZATION

Sample selection for generalization

Random sampling – while statistically ideal – is very rare in large-scale evaluations.

- This begins with well-defined population and uses probability sampling methods to select the sample.

Convenience sampling – wherein researchers begin with the schools or sites they know best or have previous experience with and work out from there – is much more common.

- This begins with the sample and then, typically concludes that results generalize to sites "like" those in the sample.

Sample selection for generalization

Random sampling – while statistically ideal – is very rare in large-scale evaluations.

- This begins with well-defined population and uses probability sampling methods to select the sample.

Purposive sampling – what I'm going to talk about now – offers a middle ground option. Like random sampling, this starts with a well-defined population.

- Like convenience sampling, it recognizes that making generalizations will require assumptions.

Convenience sampling – wherein researchers begin with the schools or sites they know best or have previous experience with and work out from there – is much more common.

- This begins with the sample and then, typically concludes that results generalize to sites "like" those in the sample.

Purposive sampling

After an experiment is over, we often generalize by asking:

- How similar is the sample to a population on some important covariates?
- (Unfortunately, if not similar, we can't do much about it).

Purposive sampling asks, before the experiment begins:

- What **inference population** is appropriate for this study?
- What **covariates** do we think matter? (Those that explain variation in treatment impacts).
- **How** can we select our sample so that it is, in fact, similar to this inference population?

Stratification

One method for selecting a sample that is similar to a population is **stratification**.

For example, if the population includes: Urban, Rural, Town, and Suburban areas, our sample should include some of each too.

When there are many covariates, this is harder. One method (Tipton, 2014) is to use k-means cluster analysis to make these strata.

Benefits

Using strata ensures that:

- An **inference population** is well-defined.
- The sample selected at the end is a **miniature** (in terms of frequency) of the population.
- A **recruitment plan** is developed that is targeted.
- Recruiters **"see"** a large pool of potential schools, not just those they are familiar with.
- **Non-response** can be tracked (allowing future analyses).

In action

Open your browser (Google Chrome is best) and go to:

www.thegeneralizer.org

Tool #2

ASSESSING SIMILARITY

The problem

For a study **already completed**:

How similar is the experimental sample to different relevant inference populations?

For example:

- United States
- Each of the 50 states
- Other relevant groups

Comparisons

We could compare features of the sample and population:

- Are the means similar on important covariates?
- Have we included all relevant sub-types?

Comparing: Similar?

	Population	Experiment (Sample)
Teacher tenure (mean years)	7.09	6.80
Teacher experience (mean years)	11.58	10.95
Teacher-student ratio	12.70	13.27
Teachers that are African American (%)	8.39	2.56
Teachers that are Hispanic (%)	14.72	21.57
Teachers in the school (total)	39.87	42.98
Teachers in first year of teaching (%)	8.32	8.74
Teachers with 1-5 years experience (%)	28.01	28.74
Teachers with > 20 years experience (%)	20.25	17.70
Students in disciplinary alternative education programs (%)	3.10	3.42
7th grade retention (rate)	1.83	1.31
Students that are mobile (%)	19.23	14.80
Students in school that are in 7th grade (%)	31.21	34.99
Students in 7th grade (total)	190.40	224.25
Students that are African American (%)	11.79	5.11
Students that are Hispanic (%)	40.27	47.19
Students that are LEP (%)	7.54	9.44
Students that are economically disadvantaged (%)	53.64	52.08
Students that are at risk (%)	43.47	40.61
Students proficient in 7th grade reading (%)	81.90	86.00
Students proficient in 7th grade math (%)	72.79	75.56
Students proficient in grades 3-11 math (%)	73.60	75.01
Students proficient in grades 3-11 all (%)	63.29	63.84
Students with commended performance, grades 3-11, math (%)	19.61	20.32
Students with commended performance, grades 3-11, reading (%)	8.71	8.88
County of school is rural	0.33	0.32

Problems

- How do we summarize the overall similarity between the sample and population?
- How do we know if it is “similar enough”?
- Imagine replicating this 50 times, producing separate tables for each state!

Solution

What we need is a **single number summary** that that provides the degree of similarity *across all these covariates.*

The **generalizability index** (Tipton,2014) does exactly this:

- It is simple to compute;
- It takes values between 0 and 1;
  - 1 indicates the sample is an exact miniature of the population;
  - 0 indicates they share no common features;
- Its values indicate when:
  - a sample is like a random sample;
  - post-hoc adjustments are useful;
  - generalizations are unwarranted.

How does it work?

One way to compare a sample to a population on a set of covariates is through a **propensity score**.

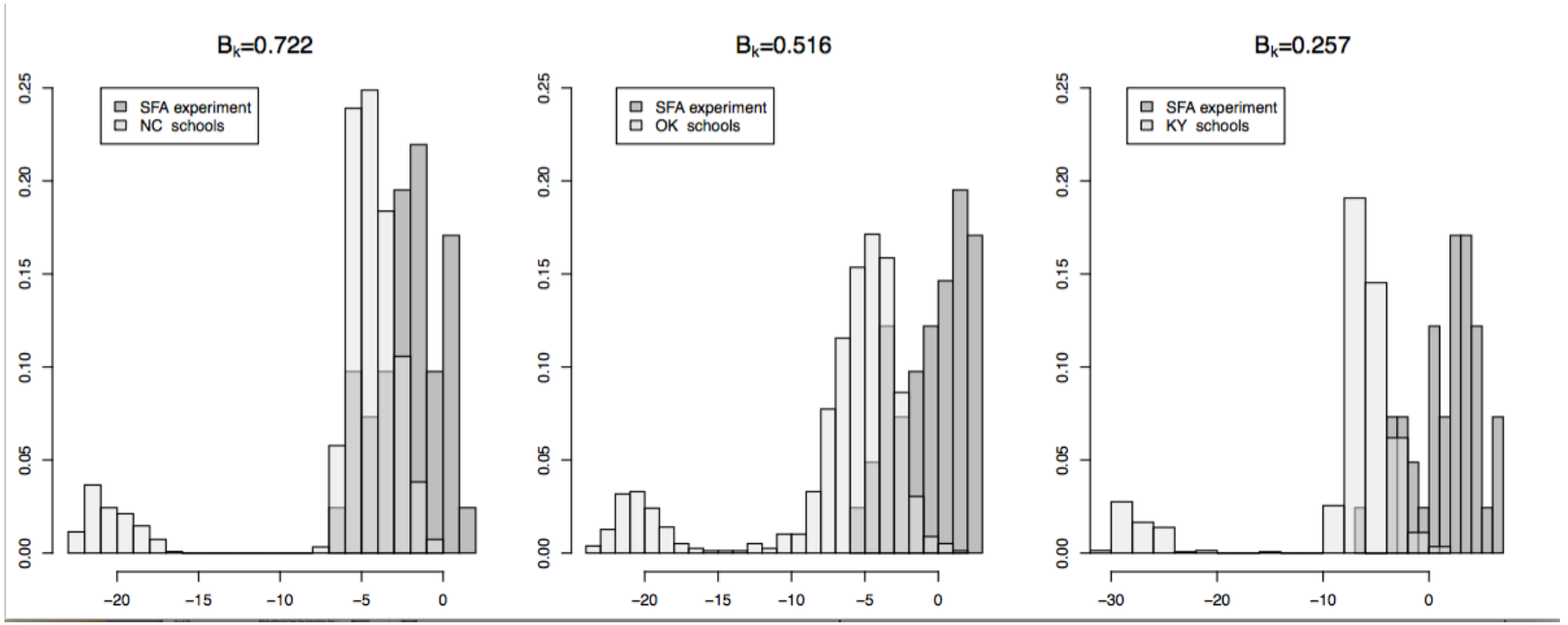
This could be estimated using logistic regression.

The distributions of propensity scores (or their logits) could be compared.

Differences in these distributions can be summarized through a statistic.

Visualization

(We want a # to summarize this)

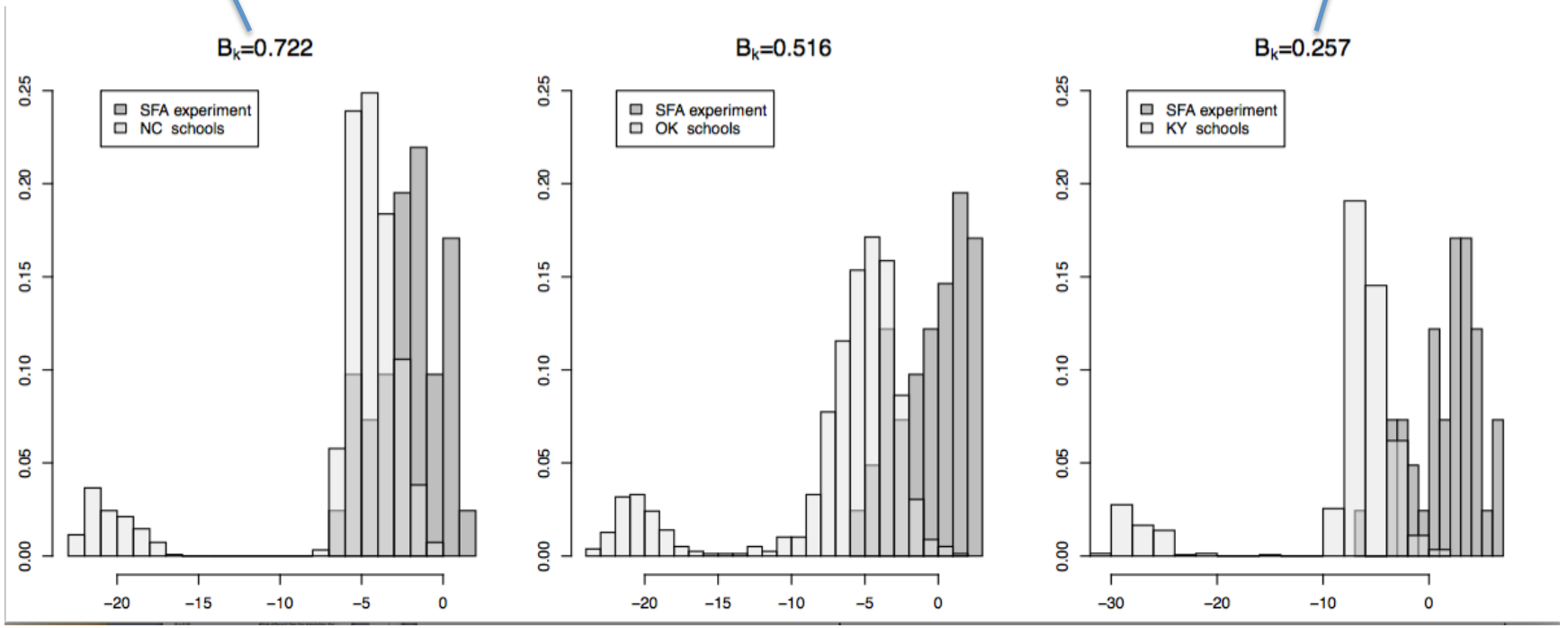


Visualization

(We want a # to summarize this)

Similar

Different



Details

If we divided these densities (histograms) into k bins:

$$B_k = \sum_{j=1}^{k} \sqrt{w_{p_j} w_{s_j}}$$

It can be shown that we can also write this as

$$B_k = \sqrt{\theta\tau} \sum_{j=1}^{k} \sqrt{w_{p0_j} w_{s0_j}}$$

Where

- **θ** measures the proportion of the population which overlaps with the sample;
- **τ** measures the proportion of the sample that overlaps with the population; and
- $\mathbf{w}_{pj0}$ and $\mathbf{w}_{sj0}$ are the weights within the overlap region.

What do the values mean?

Values of Generalizability Index	Conclusion	Next step
B > 0.90	The sample is as similar to the population as a random sample (on the covariates studied).	Use the sample ATE to estimate the population ATE.
B > 0.70	The differences are minimal and can be adjusted well (small changes in standard errors; most bias removed).	Reweighting needed. Try IPW or post-stratification using propensity scores.
0.50 < B < 0.70	Some adjustments may be possible, but costly (large increases in standard errors/ limits to bias reduction).	Reweighting needed. Try IPW or post-stratification using propensity scores.
B < 0.50	The sample is sufficiently different from the population that generalizations are not possible.	Reweighting is unlikely to be successful. Limit scope of generalizations.

In action

Open your browser (Google Chrome is best) and go to:

www.thegeneralizer.org

Take home points

1. **Relevance is always about generalization.**
2. **Generalizations will be made**.
   - The question is not "do we want to generalize?" but instead "do we want to **lead** the generalizations?"
3. **If at all possible, plan for generalization**.
   - Every study has an inference population: some are broad, while others are narrow.
   - Even when your best efforts fail, you will be in a better situation for post-hoc statistical adjustments.
4. **When not possible, help others understand where results generalize.**
   - Be sure to report your **assumptions**: Which covariates did you compare on. Why these?

Thanks!

Elizabeth Tipton

Teachers College, Columbia University

tipton@tc.columbia.edu

http://blogs.cuit.columbia.edu/let2119/

Also

If you use The Generalizer and run into problems, please let me know:

- Note the page (the % bar at top);
- If appropriate, send a screen shot;

And also:

- If there are features you think that would be useful, let me know!

Figure 1: Three regions when comparing densities