

Evaluation of Public Policies and Social Programs

Session 16: External Validity — From Pilot (or Other Schools) to This School

Andrés Barrios Fernández

Wednesday, October 28, 2026

MA in Economics, Universidad de los Andes, Chile — Second Semester 2026

1. What does a credible estimate tell us — and what does it not?
2. Why effects vary across sites and when programs scale
3. How transferable is Astoreca's track record to a new school?
4. Team exercise: build a transparent transportability judgment

No paper presentation today; use the session to strengthen the evidence behind your final estimate.

Learning Objectives

By the end of today's session, you will be able to:

- Distinguish internal validity (causal identification in the study sample) from external validity (relevance to a new setting)
- Identify mechanisms that can change effects across populations, contexts, implementation, and scale
- Assess which features of Astoreca's existing schools are similar to or different from the proposed new site
- Communicate a transported estimate as a justified point estimate or range, with explicit assumptions

From Causal Effect to Decision-Relevant Effect

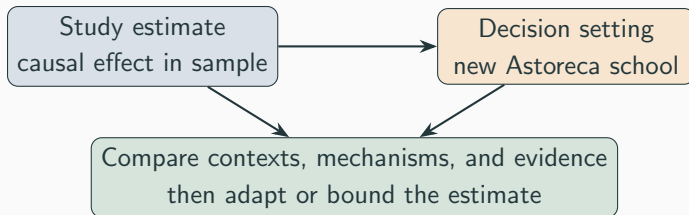
A strong causal estimate can still be the wrong forecast

Suppose a study credibly finds that a school model raised achievement by 0.20 standard deviations.

- Internal validity asks: did the program cause that effect *for the study population, at that place and time?*
- External validity asks: what effect should we expect for a different population, location, implementation team, or scale?
- The first question is about the counterfactual. The second is about transport from the study setting to the decision setting.

A credible estimate is evidence for a forecast — not automatically the forecast itself.

A transportability map



- If relevant features are similar and the mechanism is stable, direct transfer may be reasonable.
- If important features differ, adjust using evidence when possible; otherwise show a range and explain its basis.
- If the mechanism itself may change, do not disguise model uncertainty as a precise adjustment.

Why Effects Differ

What does the cross-site evidence say?

Vivalt (2020), “How Much Can We Generalize from Impact Evaluations?”

Impact estimates vary considerably across contexts. A study's reported effect is not a universal constant that can be copied unchanged into every new setting.

- Site-level variation may reflect real differences in populations, institutions, and implementation — as well as measurement and sampling uncertainty.
- The published evidence base is selected: publication and reporting processes can make the observed studies unrepresentative of all attempts.
- Vivalt also documents systematic differences across types of implementing organizations; this is a warning to examine who delivered the intervention and under what conditions, not a rule that every NGO effect must shrink by a fixed amount.

Vivalt, E. (2020), *Journal of Development Economics*, 147, 102485.

Four mechanisms that can change an effect at scale

1. **General-equilibrium effects:** expansion changes wages, prices, school competition, or the availability of teachers.
2. **Implementation quality:** pilot attention, expert staff, and close monitoring may not survive routine delivery at scale.
3. **Population heterogeneity:** the next students reached may have different needs, constraints, or outside options.
4. **Political-economy effects:** expansion can alter incentives, targeting, budgets, accountability, or support for the program.

Implication

Ask not only “is the program the same?” but also “does expansion change the environment or the people it reaches?”

Muralidharan & Niehaus (2017), “Experimentation at Scale,” *JEP*, 31(4), 103–124.

Astoreca: what might travel, and what might not?

Potentially comparable

- Educational model and curriculum
- Teacher training and instructional routines
- Organization's accumulated know-how
- Similar age groups and outcome measures

Potentially different

- Student baseline achievement and needs
- Neighborhood, peers, and other schools
- Staff recruitment and leadership
- Capacity, start-up phase, and scale

These are hypotheses to test with evidence, not assumptions that all similarity or difference matters equally.

Team Discussion: Transport the Evidence

Build a transportability judgment

Teams, 20 minutes

Assume Astoreca's existing schools show a positive achievement effect. You must forecast the effect for a proposed new school.

1. List the two most decision-relevant similarities and differences between the existing schools and the new site.
2. For each difference, state a causal mechanism and whether it would likely raise, lower, or leave the effect unchanged.
3. Decide whether evidence lets you adjust the estimate. If not, define a defensible range and explain what anchors its bounds.
4. Name one feasible piece of data that would most reduce your uncertainty before the final recommendation.

From transport judgment to a credible forecast

- **Direct transfer:** appropriate only when relevant population, implementation, context, and scale are sufficiently similar.
- **Evidence-based adjustment:** use observed differences and credible effect moderators; avoid multiplying by an arbitrary “external validity discount.”
- **Range:** when the direction is plausible but magnitude is weakly identified, show how alternative transport assumptions change the result.
- **New evidence:** a phased rollout or early cohort can update the forecast, but should not be presented as definitive evidence before outcomes mature.

External validity is a reasoned bridge from evidence to this decision, not a yes/no label.

What We Learned Today

- Internal validity establishes a causal effect in the study setting; external validity asks how well it transports to the decision setting
- Effects may change through general-equilibrium feedback, implementation quality, population heterogeneity, or political economy
- Astoreca's track record is valuable evidence, but its relevance depends on the new school's people, place, delivery, and scale
- Adjust only when supported by evidence; otherwise show a transparent range and identify what information would narrow it

- Finalize your analysis and prepare to connect it to the complete private-versus-social evaluation
- Bring your headline estimate, key assumptions, and largest remaining uncertainty