Why We Cannot Measure What Works

Infrastructure Gaps in Employment Services

Hannu Karhunen

Labour Institute for Economic Research LABORE

Tieteiden talo, Helsinki | 26 May 2026

"The Evolving Employment Policy of Denmark and Finland"

What do we mean by “what works”?

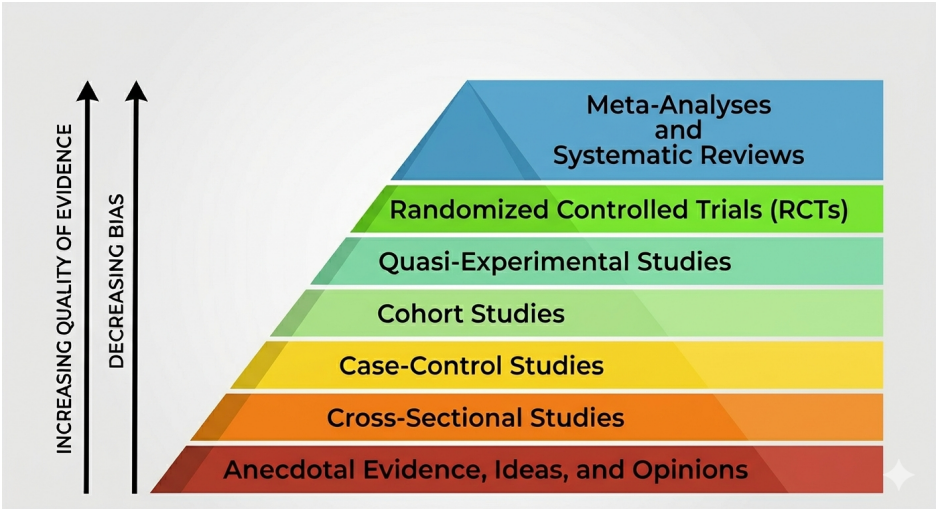
When we say we want to know **what works in employment policy**, we are asking three questions at once:

- **Causality** — does the program *cause* the outcome, or do better outcomes reflect who enters?
- **Cost–benefit** — even if it works, does it pay (direct costs, administration costs, deadweight, displacement)?
- **Evidence quality** — how confident can we be in the answer or are we just guessing and pretending?

Finnish employment policy debate is dominated by rhetoric rather than evidence.

Without causal identification and cost–benefit evaluation, “more activation” and “more ALMP funding” remain political slogans, not evidence-based policy.

Hierarchy of Evidence: Good Intentions Are Not Evidence



Causality is the bottleneck

Why correlation is not enough:

- **Selection.** Motivated jobseekers enrol ⇒ programs look effective even when they do nothing.
- **Confounding.** Macro cycles and parallel reforms move outcomes regardless of the program.
- **Displacement.** Participants take jobs from non-participants — gross effects overstate net welfare gains (Gautier et al. 2018, *JoLE*).

Only random / quasi-random assignment breaks all three cleanly.

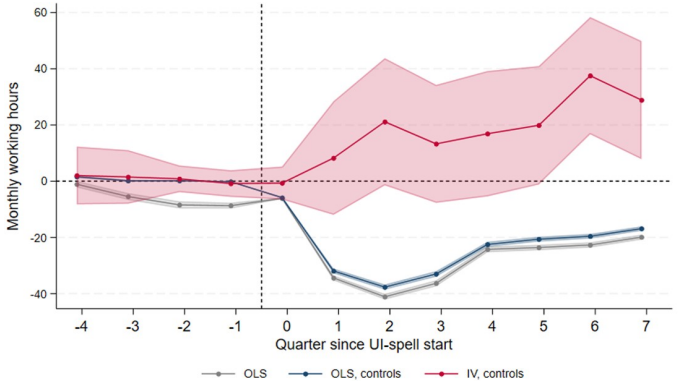
Card, Kluve & Weber (2018, *JEEA*)

Meta-analysis of **207 studies / 857 programs**: average ALMP effects are small in the short run (under one year) but *larger and positive after 2–3 years*. The most informative estimates come from randomized designs.

Pretending we know what works is dangerous

Same Danish data — opposite conclusions.

OLS suggests one policy conclusion; credible causal identification suggests another. This is why evaluation design matters. Source: Humlum, Munch & Rasmussen (2025), NBER WP 33807 (2nd-round R&R at the *Journal of Political Economy*)



Denmark is the anchor — and the benchmark for Finland

Kreiner & Svarer (2022), *Journal of Economic Perspectives*

"The Danish Ministry of Labor has organized **eleven randomized experiments since 2005**". Question: Are there already more than 20 RCTs by now?"

The features that matter for the rest of this talk:

- **Ownership.** STAR runs a permanent evidence strategy — not one-off studies.
- **Cumulative evidence.** Quickly Back to Work (2005–06) → national policy → 10-year follow-up (Bækgaard et al. 2024, *PNAS*) reveals trade-offs by client group.
- **Honest approach.** Programs that fail the cost-effectiveness bar are dropped.
- **Broad.** RCTs span job search, meetings, sick-listed, IPS for severe mental illness, and behavioural nudges.

RCTs as "Black Boxes" — a Danish debate Finland cannot have

Other presenters may have raised a legitimate critique earlier today: RCTs tell us the average effect of "classroom training or other policies," but not which caseworker, which curriculum, or which client interaction drove the result.

Fair enough — Denmark has spent 20 years opening that black box: with implementation studies, qualitative work, mixed-method designs, and now caseworker-IV approaches.

But Finland is not in that debate.

We don't have a black box to critique. We argue about whether the engine should be bigger or smaller — without a way to know whether it runs at all.
That is the madness.

Modern Research: Quantitative and qualitative evaluation must go together.

Qualitative research explains how policies and treatments are implemented in practice ("mechanism").
Both outcome measurement and careful description of the treatment are essential.

Three Reasons We Still Cannot Measure What Works in Finland

In a nutshell

The problem is not the amount of data — **it is the quality of the infrastructure**.

1. **The data infrastructure was not designed for causal evaluation.**
   - legally mandated vs. discretionary service or policy measures,
   - treatment intensity and timing,
   - or the true costs of different interventions.

   The issue is not "big data" scarcity, but measurement and data infrastructure quality.

2. **The system is decentralized without coordinated evaluation.**
   - Regions are small $\Rightarrow$ low statistical power, need of collaboration.
   - Policies and treatment practices differ across regions, making comparisons difficult.
   - There is no centralized evaluation pipeline.

3. **There is no national RCT infrastructure — and no clear owner.**
   Finland lacks routine infrastructure for randomization, intake experiments, and long-term cross-regional evaluation programs.

What would need to change

A minimum viable evidence infrastructure for Finnish ALMP:

- No system works without ownership: Finland needs a ministry-level institution with the authority and mandate to build evaluation capacity at scale — essentially a supercharged version of Denmark's STAR.
- Register data should be collected and structured for causal evaluation — not just for administrative reporting as before.
- Randomization or quasi-experimental designs should be a central part of every new program — not retrofitted after rollout.
- Civil-servant gatekeepers who actually care and have time.

Bottom line: Finland has ZERO evidence on what works

Denmark demonstrates that this is institutionally feasible. The priority is not more activation measures, larger budgets, or another reform cycle — it is creating the infrastructure that allows policies to be credibly evaluated.

Key references

- Bækgaard, M., Nielsen, S.A., Rosholm, M. & Svarer, M. (2024). Long-term employment and health effects of active labor market programs. *PNAS* 121(50): e2411439121.
- Card, D., Kluve, J. & Weber, A. (2018). What Works? A Meta-Analysis of Recent Active Labor Market Program Evaluations. *Journal of the European Economic Association* 16(3): 894–931.
- Gautier, P., Muller, P., van der Klaauw, B., Rosholm, M. & Svarer, M. (2018). Estimating Equilibrium Effects of Job Search Assistance. *Journal of Labor Economics* 36(4): 1073–1125.
- Humlum, A., Munch, J.R. & Rasmussen, M. (2025). What Works for the Unemployed? Evidence From Quasi-Random Caseworker Assignments. *NBER Working Paper* 33807.
- Kreiner, C.T. & Svarer, M. (2022). Danish Flexicurity: Rights and Duties. *Journal of Economic Perspectives* 36(4): 81–102.
- Maibom, J., Rosholm, M. & Svarer, M. (2017). Experimental Evidence on the Effects of Early Meetings and Activation. *Scandinavian Journal of Economics* 119(3): 541–570.
- STAR — Danish Agency for Labour Market & Recruitment. Evidence-based policy-making. star.dk/en/evidence-based-policy-making/