

ADVANCING POLICY THROUGH RANDOMISED EXPERIMENTS

Nordic Economic
Policy Review
2026



Contents

Authors and Editors	4
Roope Uusitalo and Antti Kauhanen: Introduction	6
Simen Markussen and Marie Bjørneby: The Norwegian Tax Experiment - From Idea to Implementation	16
Comment by Kaisa Kotakorpi	38
Comment by Helena Keinänen, Jarno Tuimala, and Olavi Tuomi	44
Ramin Izadi and Matti Sarvimäki: Learning to Experiment: Practical Lessons from Finland's Two-Year Preschool Experiment	47
Comment by Morten Visby Krægpøth	87
Comment by Ingeborg Foldøy Solli	89
Johan Vikström: Evaluating Active Labour Market Policies Using Randomised Control Trials	91
Comment by Jukka Mattila	120
Comment by Kari Hämäläinen	124
Stine Nyhus Larsen, Nikolaj Noer Poulsen, Michael Rosholm and Katrine Bønneland Tølbøll: Childhood and Adolescent Interventions in a Danish Setting: A Meta-Analysis of Evidence from TrygFonden's Centre for Child Research	127
Comment by Marjo Kyllönen and Kristiina Hannukainen	175
Comment by Hannu Karhunen	180
Elias Einiö and Annika Nivala: Large-Scale Randomised Experiments with Population Data	183
Comment by Sofie Cairo	203
Comment by Antti Kauhanen	209
About this publication	212

This publication is also available online in a web-accessible version at
<https://pub.norden.org/nord2026-001>

Authors and Editors

Editors: **Roope Uusitalo** and **Antti Kauhanen**

Technical Editor: **Lasse Leinonen**

Text Editor: **Tam McTurk**

Introduction

Roope Uusitalo, Professor, University of Helsinki

Antti Kauhanen, Research Director, ETLA Economic Research

The Norwegian Tax Experiment – From Idea to Implementation

Simen Markussen, Director, Ragnar Frisch Centre for Economic Research

Marie Bjørneby, Norwegian Ministry of Finance

Comment by **Kaisa Kotakorpi**, Professor, University of Tampere

Comment by **Helena Keinänen, Jarno Tuimala, and Olavi Tuomi**, Finnish Tax Administration

Learning to Experiment: Practical Lessons from Finland's Two-Year Preschool Experiment

Ramin Izadi, Senior Researcher, VATT Institute for Economic Research

Matti Sarvimäki, Associate Professor, Aalto University

Comment by **Morten Visby Krægpøth**, Head of Division, Danish Ministry of Children and Education, Denmark

Comment by **Ingeborg Foldøy Solli**, Professor, University of Stavanger

Evaluating Active Labour Market Policies Using Randomised Control Trials

Johan Vikström, Adjunct Professor, University of Uppsala and Researcher, Institute for Evaluation of Labour Market and Education Policy

Comment by **Jukka Mattila**, Director, VATT Institute for Economic Research

Comment by **Kari Hämäläinen**, Chief Researcher, VATT Institute for Economic Research

Childhood and Adolescent Interventions in a Danish Setting: A Meta-Analysis of Evidence from TrygFonden's Centre for Child Research

Stine Nyhus Larsen, Research assistant, TrygFonden's Centre for Child Research

Nikolaj Noer Poulsen, PhD Student, Aarhus University

Michael Rosholm, Professor, Aarhus University

Katrine Bønnelund Tølbøll, PhD Student, Aarhus University

Comment by **Marjo Kyllönen**, Head of Department, City of Helsinki and **Kristiina Hannukainen**, Evaluation Expert, City of Helsinki

Comment by **Hannu Karhunen**, Research Leader, The Labour Institute for Economic Research (Labore)

Large-Scale Randomised Experiments with Population Data

Elias Einiö, Chief Researcher, VATT Institute for Economic Research

Annika Nivala, Senior Researcher, VATT Institute for Economic Research

Comment by **Sofie Cairo**, Assistant Professor, Copenhagen Business School

Comment by **Antti Kauhanen**, Research Director, Etlä Economic Research



Nordic Economic Policy Review 2026

Advancing Evidence-Based Policy: Insights from Randomised Experiments in the Nordic Countries

Antti Kauhanen and Roope Uusitalo

1 Introduction to the issue

Randomised controlled trials (RCTs) are the gold standard of the scientific approach. The natural sciences have a long tradition of relying on experiments. Passing randomised trials conducted according to strict protocols is a prerequisite for launching new products. The trials must demonstrate that the positive effects of the new medicine are sufficiently large and that the potential side effects are sufficiently small compared to those of existing drugs.

By contrast, public policies can be implemented with high hopes but little evidence of their likely effects. To a degree, this lies at the heart of our democratic systems. Governments and parliaments make political decisions and rely on the support of voters, not on the support of scientific research. Voters have conflicting interests, and political decisions reflect the political strength of different groups of voters. Politics cannot be reduced to the implementation of scientifically tested treatments.

However, an increasing number of governments rely on, or at least claim to rely on, evidence-based policies. In an ideal world, political decision-makers would have unbiased and reliable information on the likely outcomes of potential policy choices and can base their actions on that information. However, such a shift places high demands on the quality of the evidence.

Over the past 30 years, economics has undergone a credibility revolution. Better research designs have improved the quality of the evidence and made the results more convincing. Better designs also enable researchers to analyse causal effects rather than only correlations.

The credibility revolution initially increased the use of "natural experiments", i.e., situations in which the target population is divided into "treatment" and "control" groups without anyone actively allocating potential participants into these groups. A bit later, RCTs also became more commonplace in economics. The key difference is that random allocation to the treatment and control groups is monitored and often implemented by the evaluation team. RCTs have been used to evaluate planned reforms in active labour market policy (ALMP), education, health, and several other fields.

In a randomised trial, there are no systematic differences between the treatment and control groups. Any differences between groups after the treatment that exceed what would be expected to occur by chance can, therefore, be interpreted as the effect of the treatment. This makes evaluating effects simple, transparent, and convincing.

Even if evaluating the effects of experiments is easy, designing them is anything but. This is particularly true in policy experiments implemented in real-life settings outside the laboratory. Experiments involve addressing important issues, such as the design of the experiment, co-operation with organisations implementing the policies, the legal and ethical grounds for treating potential participants unequally, and the use of the results.

Researchers in all of the Nordic countries have actively used an experimental approach to evaluate various policies. The treatments that have been analysed range from "light" ones involving sending encouragement letters to participants to massive multi-million budget experiments, changing the rules of the social security system for a randomly selected group, as in the Finnish basic income experiment, or changing the parameters in the tax system, as in the new Norwegian tax allowance experiment.

Researchers in Nordic countries have one major comparative advantage in conducting policy experiments: access to population-wide high-quality register data. Access to these data solves many issues related to the representativeness of the treatment and control groups. As far as outcomes from the experiments can be measured from administrative register data, the selectivity of the respondents to follow-up surveys generates much smaller risks of bias in the results. In some cases, the experiment can be conducted without even contacting the control group, which often has very low response rates due to minimal incentives to respond to surveys related to experiments in which they did not take part. In addition, data collection is often one of the most expensive parts of an experiment. Using administrative data means that the data-acquisition costs do not depend on the sample size, allowing much larger treatment and control groups and, hence, much more precise estimates.

In this issue of the Nordic Economic Policy Review, researchers who have conducted randomised policy experiments in the Nordic Region share their experiences. Hopefully, these experiences will be useful to those planning future experiments and to those using the experimental results to guide policy choices. In this introduction, we introduce some of the issues covered in later chapters.

2 Issues involved in experiments

2.1 Degree of control in controlled experiments

In an ideal RCT, the participants are randomly allocated to the treatment and control groups. Randomisation ensures that the treated group is, on average, similar to the control group in all aspects. All individuals in the treatment group receive the treatment, and none in the control group do. Finally, treatment only affects the treatment group and has no impact on the control group. In practice, however, this may not be the case.

Randomisation does not guarantee that all individuals allocated to the treatment group will receive the treatment. Some may refuse it, and some may drop out during the experiment. This question is especially relevant when the 'treatment' is a service rather than, e.g., a financial subsidy. If individuals cannot be forced to participate, researchers often estimate the impact of offering it as in the chapter on recruitment subsidies by Einiö and Nivala. As noted in the chapter by Sarvimäki and Izadi, even if participation in an experiment is in principle mandatory, the take-up is often incomplete as in the Finnish pre-school experiment. In cases when allocation to treatment is decentralised, researchers have an additional problem and must make sure that all units follow the same protocol in assigning individuals to treatment and control groups. This might be challenging, for example, in experiments on employment services, where multiple agencies are responsible for implementing the experiment. This issue is discussed in Vikström's chapter on ALMP experiments.

It may also be the case that individuals allocated to the control group receive treatment from another provider or receive something similar to the treatment. For example, in the case of ALMP, individuals allocated to the control group in an experiment on enhanced job search assistance may buy similar services from private providers, or in decentralised experiments, some offices may not follow the randomisation protocol, and some individuals in the control group may be treated. These implementation problems complicate interpretation of the results.

In many cases, the treatment may also affect the control group. This is a serious problem. The role of the control group is to estimate what would have happened to the treatment group if it had not been treated. If the treatment also affects the control group, it cannot serve this purpose. Vikström's ALMP chapter discusses an important case in which these spillover effects may be significant: enhanced job search services. If

some unemployed job seekers are offered an enhanced service, it may improve their chances of finding a new job, but it may also lower the chances of other job seekers.

Spillover effects are also discussed in the chapter on the Norwegian tax experiment. During the design phase, the researchers discussed whether to conduct the experiment at the individual or regional level. A regional analysis would have captured potential spillover effects, which, in this case, might stem from increased labour demand driven by higher purchasing power.

However, incomplete take-up, spillover effects, or alternative treatments for the control group do not invalidate the experimental approach, although they make the analysis more complicated. The chapters in this volume demonstrate how the researchers approach the issue and manage to uncover effects of treatment also in imperfect experiments.

2.2 Following participants over time

To evaluate experiments, researchers need to have outcome measures for both the treatment and control groups. Pre-treatment information is also often useful for validating the approach and obtaining more precise results. Data acquisition for experiments is costly, especially if the data are gathered by conducting surveys of potential participants. Ensuring high response rates can be expensive, and even when substantial resources are devoted to this, response rates may remain low, reducing the experiment's statistical power and potentially biasing the results if non-responses are not random.

One advantage Nordic countries often have is access to the data needed for evaluation from statistical population registers. When this is the case, non-response problems are minimised, and the costs of gathering the relevant data are small compared to those of sending out surveys. Population registers also allow tracking individuals over long periods, which is rarely feasible with surveys.

Register data becomes particularly valuable when there is a long time lag before a response to the treatment or when the effects may fade over time. For example, in the Norwegian tax experiment discussed by Markussen and Bjørneby, the experiment lasts five years, and people may adjust their behaviour slowly. In the Finnish preschool experiment discussed by Izadi and Sarvimäki, the researchers plan to follow participants throughout their school careers and develop evaluation tools for use in their follow-up work.

2.3 Legal and ethical aspects

Experiments necessarily involve unequal treatment of potential participants. Allocation to the treatment group may provide direct financial benefits to participants or

additional resources, e.g., in education or job search. Or the experiment may deny the control group access to these resources. Programmes that require compulsory participation may also require participants to act in ways they would not do voluntarily.

Unequal treatment is necessary to create variation in the treatment. However, it must be justified. Experimental research uses several approaches to address this problem.

The issues are probably least severe in experiments where participation is voluntary, and participants provide informed consent to the treatment. Most experimental medical studies rely on this principle. In the social sciences, prime examples are experiments that provide information and encourage participation in the treatment, particularly when resources are limited, and not all potential participants can receive treatment. Even in these cases, research designs typically need to be pre-approved by ethics committees, but ethical considerations rarely prevent such experiments.

Ethical issues become more important when the treatment group receives benefits from the public sector or faces mandates from public bodies. Even in these cases, unequal treatment may not be the most fundamental issue. Even in the absence of experiments, many public policies involve age limits, apply only in specific regions, or otherwise treat people differently. In experiments, the question is whether scientific research provides sufficient justification for unequal treatment.

As shown by the authors in this edition of NEPR, unequal treatment of control and treatment groups is possible as long as the goals of the experiment are sufficiently important and as long as the experiment does not cause disproportionate harm to the participants. In some cases, it is also necessary that the experiment is governed by law. Following these principles, specific legislation was introduced in Finland compelling randomly selected five-year-olds to participate in preschool (Izadi & Sarvimäki), and legislation granting tax exemptions to a randomly selected cohort of young people was introduced in Norway (Markussen & Bjørneby).

Legal mandates can also be useful in other ways than enabling the experiment. For example, the Finnish employee payroll tax experiment granted the researchers full access to data (Einiö & Nivala), and the Finnish preschool experiment made it compulsory for daycare centres and schools to test children's skills (Izadi & Sarvimäki).

2.4 What would we know without experiments?

Given the cost of conducting experiments, it is natural to ask whether it would be possible to use non-experimental methods to conduct credible evaluations of policies. This discussion began in economics in 1986, when Robert Lalonde published his landmark paper, showing that the non-experimental methods available at the time could not reproduce the estimates obtained from large-scale experiments. Further studies have shown that better data (for example, similar measurements of the treatment and control groups) and more advanced methods help substantially reduce

bias in non-experimental evaluations (Heckman et al. 1998, Smith and Todd 2005, etc.). A very recent paper shows that the most recent non-experimental methods can replicate experimental results. However, to enable causal interpretation, researchers should always assess the key assumptions (Imbens and Xu, 2025). Therefore, the answer to the question of whether non-experimental methods can be used to conduct credible evaluations of policies is that it depends on the quality of the data and on whether the treatment is more or less randomly assigned.

The chapters in this edition touch upon this question several times. The chapter on experiments in education by Nyhus Larsen, Noer Poulsen, Rosholm and Bønneland Tølbøll includes both experimental and non-experimental evaluations. It shows that, on average, the results are similar. This supports the idea that, at least in this setting, experimental and non-experimental approaches can be used to evaluate policies. On the other hand, Vikström's ALMP chapter discusses an example in which experimental and non-experimental evaluations led to opposite conclusions about the effectiveness of a policy. In that case, decisions based on non-experimental evaluation would have led to the wrong policy being implemented. The chapter on the Norwegian tax experiment by Markussen and Bjørneby explains that the experiment was conducted precisely because the policy could not be evaluated by any other method than an experiment. Finally, the chapter on the recruitment subsidy by Einiö and Nivala shows that the experiment's results were more precise than those of a non-experimental evaluation of an earlier recruitment subsidy. This case highlights another difference between experimental and non-experimental evaluations. When the experiment was designed, the way the subsidy was implemented was changed to improve take-up. Thus, the evaluation led to a rethink of how to implement the subsidy to maximize its impact.

2.5 What can the experimental approach be used for?

As noted above, in many cases, carefully conducted non-experimental research can yield results similar to those of experimental research. However, this always depends on the treatment's selectivity and the opportunities to address selectivity issues through research design. Sometimes this is not possible, and in all cases, the credibility of non-experimental research depends on the credibility of its assumptions.

A randomised trial is an ideal way to solve selectivity issues. However, they also present problems. One important factor is publication bias. If journal editors or the researchers themselves prefer statistically significant results, it is possible that only large and significant findings will be published. This could easily lead to incorrect conclusions based on the cumulative empirical evidence.

Fortunately, there is a simple solution to this problem. Requiring that experiments be pre-registered before the results are known increases the likelihood that also insignificant findings are reported and reduces the incentive to run many experiments and report only those with the desired results.

Another issue is that the experiments are time-consuming. Designing, conducting, and monitoring the outcomes of an experiment takes years, and this often exceeds the patience of policymakers elected for a fixed term. In some cases, as discussed by Izadi and Sarvimäki in this edition, it may be possible to convince anxious policymakers to implement a pilot or, rather, a controlled experiment, instead of rolling out a full-scale reform. Unfortunately, the results of experiments do not always match the expectations policy designers had for their favourite programmes. Whether this is a positive result, saving money that would have been used for the policy, or a negative one sinking a great idea that would have been implemented without a disappointing experiment, probably depends on the observer.

3 Description of the chapters

The chapter by Ramin Izadi and Matti Sarvimäki details the evaluation of the Finnish two-year preschool experiment. It involved approximately 35,000 children at more than 1,000 centres in 148 municipalities, facilitated by temporary legislation that made participation compulsory, specified randomisation, and authorised extensive data collection.

The treatment brought preschool forward by a year (from age six to age five), significantly altered learning environments by introducing more qualified teachers, more homogeneous age groups, and slightly more guided learning time, but leaving overall resources and schedules largely unchanged. A central contribution of the experiment is the data infrastructure: large-scale direct assessments of academic (numeracy, literacy) and teacher-rated socioemotional skills (task performance, social skills, emotional regulation, peer relationships) at ages 5–7 for over 30,000 children, expanding to nearly 60,000 at school entry, linked to rich administrative records covering children, families, teachers, peers, schools, and centres, with long-term follow-up until the early 2030s.

The results show that extending pre-school had sizable short-term gains in academic and socio-emotional skills for the children who were shifted from home care to centre-based pre-school. However, for the large majority of children, the alternative was standard daycare. For this group extending pre-school had no effect on skills measured at school entry. As a result, average effects at the population level were essentially zero.

The chapter by Simen Markussen and Marie Bjørneby presents a significant real-world Norwegian tax experiment aimed at generating evidence for policymaking. This experiment involves randomly granting an earned income tax allowance to approximately 100,000 young taxpayers – about 8% of the 20–35 age group – through their 2025 tax cards. The experiment started in January 2026, so no results are available yet. The chapter details the design choices, anticipated responses, ethical considerations, and implementation of the experiment on a national scale. Power

analyses indicate that with 100,000 recipients, the effects on employment and earnings are estimated with precision. The chapter also explores ethical trade-offs, such as minor indirect costs to non-recipients. It discusses consent, noting that the ministry deemed an opt-out option infeasible before parliamentary approval, given the absence of negative risks to anyone. In addition, it addresses legal proportionality and reasonable justification for unequal treatment after dismissing non-randomised alternatives as insufficiently informative.

The chapter by Johan Vikström surveys how RCTs can credibly evaluate active labour market policies, explaining why non-experimental comparisons are vulnerable to selection and stressing RCTs as the gold standard for identifying "what works, for whom, and under what conditions." The discussion is structured around three examples: a Danish study of private providers, a Swedish "early meetings" intervention, and a Norwegian programme focused on a light-touch habits programme. These trials are used to offer practical advice on ethics, randomisation designs, pre-analysis plans, implementation challenges, and maintaining research independence. A key message is that random assignment can be as ethical as any alternative when resources are scarce, provided transparency and safeguards are in place. The chapter then turns to measurement and analysis, highlighting the strengths of administrative data (e.g., reporting incentives and shifting responsibilities), the importance of studying take-up, the value of boosting compliance, and the use of objective outcome and cost data. Finally, the chapter urges designers to consider seriously general equilibrium responses and discusses the typical sources of equilibrium effects and how they can be studied using two-level randomisation.

The chapter by Elias Einiö and Annika Nivala examines Finland's Recruitment Subsidy Experiment, shedding light on its implications for other large-scale randomised national experiments that involve population registers and costly, imperfectly adopted treatments.

This chapter underscores the primary design challenge of balancing fixed budgets with uncertain take-up, detailing the trade-offs between under-subscription, which mainly results in a loss of statistical power, and over-subscription, which can introduce bias. It proposes a staged strategy to estimate take-up with minimal risk of over-subscription and then scale offers by explicitly managing the tolerated probability of over-subscription using the sampling distribution of the estimated take-up rate. The chapter also discusses how population register data can eliminate survey costs, enhance planning and implementation, and enable precise estimation of the effects of offering treatment on a national scale.

The results indicate that recruitment subsidy had a significant effect on hiring, although its total employment effect was rather modest, at 207 jobs. Compared to previous non-experimental evaluation, the experiment managed to induce much higher take-up rate and more precise estimates and hence, potentially different conclusions on effectiveness of the policy.

The chapter by Stine Nyhus Larsen, Nikolaj Noer Poulsen, Michael Rosholm, and Katrine Bønneland Tølbøll synthesises a decade of experimental and quasi-experimental intervention research by TrygFonden's Centre for Child Research in Denmark to estimate learning outcomes and explain the variation in effects across programmes and contexts. The authors report modest yet positive average improvements in outcomes, with more significant impacts observed for preschool-age targets, daycare settings, educator-focused delivery, and an emphasis on process quality. The average effects do not differ between the RCTs and quasi-experiments.

This study updates and extends Rosholm et al. (2021), incorporating more RCTs and all available non-RCT causal studies from the same research centre, and replicates the pattern that earlier investments yield larger returns, consistent with the so-called Heckman curve. While these design- and context-sensitive insights are valuable for policy, the authors are transparent about the following limitations: short-run outcomes, potential cost measurement errors, clustering and spillovers, and generalisability, primarily to the Scandinavian welfare context. In addition, the focus on one centre – chosen to ensure methodological and contextual homogeneity and rich cost information – necessarily excludes other Nordic interventions, potentially leading to a skewed picture of regional effectiveness.

Taken together, the chapters in this volume illustrate both the promise and the complexity of using randomised experiments to inform public policy in the Nordic countries. They show that high-quality administrative data, careful design, and close collaboration between researchers and public institutions can generate exceptionally credible evidence, while also highlighting the ethical, legal, and practical challenges inherent in experimenting with real-world policies. As governments increasingly look for solid evidence to guide decisions, the lessons from these studies provide practical help for designing better trials, understanding their results, and using experimentation more routinely in policymaking. We hope this collection not only contributes to current discussions but also encourages Nordic policymakers and researchers to keep promoting a rigorous, transparent, and learning-oriented approach to public policy.

References

- Heckman, J. J., Ichimura, H., Smith, J. A., & Todd, P. E. (1998). Characterizing selection bias using experimental data. *Econometrica*, 66(5), 1017–1098.
- Imbens, G. W., & Xu, Y. (2025). Comparing experimental and non-experimental methods: What lessons have we learned four decades after LaLonde (1986)? *Journal of Economic Perspectives*, 39(4), 173–202.
- LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76(4), 604–620.
- Rosholm, M., Paul, A., Bleses, D., Højen, A., Dale, P. S., Jensen, P., Justice, L. M., Svarer, M., & Andersen, S. C. (2021). Are impacts of early interventions in the Scandinavian welfare state consistent with a Heckman curve? A meta-analysis. *Journal of Economic Surveys*, 35(1), 106–140.
- Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of non-experimental estimators? *Journal of Econometrics*, 125(1), 305–353.



Nordic Economic Policy Review 2026

The Norwegian Tax Experiment – From Idea to Implementation

Simen Markussen and Marie Bjørneby

Abstract

In December 2025, 106,469 young Norwegians received their annual tax card (skattekort), which adds a new feature to the tax system in Norway: an earned income tax credit. The recipients are all part of a large real-world experiment, in which 8% of all taxpayers between the ages of 20 and 35^[1] were randomly assigned to a treatment group. For this group, the tax system includes an additional earned income tax credit.

In this article, written before any results from the experiment have materialised, we present the experiment, how it came about, and how we worked on its implementation. We aim to present some of the challenges we encountered, as well as the insights we gained while working on the project.^[2]

-
1. They will continue to receive the allowance after turning 36, as it is granted by cohort, not age.
 2. This article is partly based on the early assessment by a joint research group at Norwegian Fiscal Studies, Department of Economics at the University of Oslo and the Ragnar Frisch Centre for Economic Research available [\[here\]](#). The research group consists of Simon Bensnes, Øystein Hernæs, Morten Håvarstein, Simen Markussen, Magne Mogstad, Oddbjørn Raaum, Ragnhild Schreiner and Gaute Torsvik.

1 Introduction

From 1 January 2026, a randomly selected 8% of all taxpayers in Norway aged 20–35 have a different tax schedule than the rest of the population. The modification includes an earned income tax credit (EITC) targeting people on low and medium incomes. The goal, which is now being put to the test in a large-scale randomised experiment planned to last five years, is to increase labour market participation. In this article, written before any results from the experiment have materialised, we present the experiment, what we can learn from it, how it came about, and how we worked on its implementation. We aim to present some of the challenges we faced as well as the insights we gained while working on this project.

The article is structured as follows. In Section 2, we describe the background to the experiment as the intersection of three different developments. First, a negative development in labour market participation among young adults, at least relative to other countries. Second, an interest in EITC and whether it is a suitable way to increase labour market participation. Third, following the pandemic, there has been an increased awareness of the need for knowledge-based policy and a growing interest in randomised trials. In Section 3, we describe the feasibility study conducted by the research group and the Ministry of Finance, which assessed the expected fiscal and behavioural effects of different designs for EITCs, looked at the opportunities for learning about those effects by using different research designs, and discussed how these insights relate to the legislative requirements for experiments of this kind.

In Section 4, we briefly describe the public hearing and the reactions to the proposed experiment. In Section 5, we describe the practical planning and implementation of the experiment in summer and autumn 2025, in particular, the information sent from the tax administration to randomised groups of taxpayers, the stratified randomisation, a large complementary survey, research ethics, and more.

2 Background

The large-scale experiment can be seen as the result of three developments over the last few years. First, in Norway, as in many other countries, improving labour market participation remains a key challenge for public policy. Over the last 20–25 years, employment rates in Norway have not risen as much as in many comparable countries. As a result, employment rates in Norway are no longer among the highest in Europe, even though they remain relatively high.

The combination of the employment rate not rising and demographic changes is placing a growing fiscal strain on public budgets. As a response, the government has declared its ambition to raise the employment rate among people aged 20–64 from 80.5% in 2023 to 82% by 2030, with a further increase to 83% by 2035. It is considered particularly important to prevent young people from becoming detached from the

labour market. Achieving these goals will require a range of policy measures designed both to improve underlying conditions for labour market attachment and to strengthen incentives to work.

One such approach is to use financial incentives to stimulate labour supply. An earned income tax credit (EITC) is a general tool to achieve this aim. It is widely used in OECD countries as an instrument to reduce the tax burden on labour and incentivise employment (e.g., the EITC programme in the US). It typically takes the form of deductions or credits applied to taxable income or directly to tax liability, and it is often, but not always, targeted at people on low and middle incomes.

Norway is one of the countries that has not introduced EITCs into the tax system. It has, however, been proposed several times in recent years, including in the budget proposal from the (outgoing) Conservative government in 2021 and in an expert commission on taxes in 2022 (NOU 2022:20). These proposals constitute the second development that has made the experiment possible.

Despite its use in many countries, there is limited knowledge on the effectiveness of EITCs as a tax instrument. Previous Nordic and international experiences, e.g., Sweden's broad jobbskatteavdraget, typically lack clear counterfactuals when implemented at full scale.

Several studies have examined how financial incentives, more generally, affect labour supply among marginal workers. A notable example is Canada's Self-Sufficiency Project, which offered financial bonuses to single mothers working over 30 hours per week. The programme increased employment and income, and reduced poverty (Michalopoulos et al., 2005).

In the US and UK, programmes like Earned Income Tax Credit (EITC) and Working Tax Credit (WTC) provide income support to low-wage workers meeting minimum work requirements. Although not experimental, research using policy variation suggests these programmes have increased labour force participation, especially among single mothers (Eissa & Liebman, 1996; Blundell & Hoynes, 2004; Meyer & Rosenbaum, 2001). However, these findings have recently been questioned by Kleven (2024).

In Norway, reforms to disability benefit rules in 2005 and 2011 showed that stronger work incentives can move people from welfare to work, particularly younger recipients (Kostøl & Mogstad, 2014; Alne, 2018). However, a Finnish experiment with a basic income for long-term unemployed people found no significant increase in employment (Verho et al., 2022).

The existing literature on EITCs suggests they can increase labour supply, but estimating their effects is challenging due to the lack of clear comparison groups (Laun, 2019). One key issue is distinguishing between income effects and substitution effects, both of which influence how individuals respond to changes in taxation. Recent

Norwegian research (Graber et al., 2025) finds both income and substitution effects to be strong, with uncompensated elasticities of 0.15 for middle-income individuals and steadily increasing with income up to around 0.5 for high earners, which has significant implications for tax policy.

This uncertainty underscores the need for greater empirical knowledge about behavioural responses to taxation. In the case of earned income tax allowances, both income and substitution effects often point in the same direction (if the allowance is phased out at higher income levels), making it crucial to understand the magnitude of the effects to avoid unintended reductions in labour supply.

The third, and comparatively unrelated, development started with the pandemic, during which the need for knowledge about policy design became acute. A number of attempts were made at the time to conduct randomised trials to learn about the effects of policies such as school closures, labour market interventions, the use of masks, and so forth. After the pandemic, the Norwegian government set up a specialist interministerial core group to strengthen the knowledge system for crisis management ("Knowledge in Crises") and address the shortcomings in evidence-based decision-making revealed by the COVID-19 pandemic. The group commissioned reports from two expert groups highlighting the need for secure and efficient data infrastructure, rapid access to real-time and raw data, and clearer legislation and ethical frameworks for data use and experimental approaches. Following this, the government has stated that it will seek to facilitate greater use of experiments in the public administration.

Although randomised controlled trials can be effective tools for generating reliable knowledge on policy effectiveness, their use has been limited due to methodological, legal, and ethical challenges, as well as insufficient expertise within government agencies (Gopinathan et al., 2025). To address these barriers, and following recommendations from the expert group, a national guidance service for policy experiments was established in November 2025. Hosted by Statistics Norway, its goal is to create a competence environment to help public-sector agencies identify information needs, assess existing knowledge, select appropriate methods, design and implement experiments, and conduct evaluations.

Setting up this service marks a tangible step toward increasing the use of experimental methods in Norwegian public administration and ensuring that new policy measures are tested more systematically before they are implemented at scale.

3 Feasibility study: researchers and officials working together

In the 2024 budget proposal, the Norwegian government announced that it intended to explore the possibility of implementing a limited pilot scheme for an earned income tax credit for young people:

"The government believes there is a need for greater knowledge about the effects of a work tax credit before considering whether it might be an appropriate measure."

"The ministry will, with the assistance of relevant research environments, investigate (...) whether the trial scheme can and should be implemented."

It was clear that the Ministry of Finance would need to work closely with researchers over a prolonged period to achieve this. In January 2024, the ministry invited relevant research communities to an open dialogue meeting to discuss the way forward and what would be required for any potential pilot to deliver the intended learning outcomes. In June 2024, the ministry entered into a research and development (R&D) agreement with a team of researchers from the Ragnar Frisch Centre for Economic Research and Norwegian Fiscal Studies at the University of Oslo, formalising a long-term collaboration to support the development and evaluation of the initiative.

Three questions were central to this assessment:

1. Was there a need for greater knowledge about how EITCs affect labour supply?
2. Could an experiment contribute to new insights – and how?
3. Would such an experiment be feasible given the legal and ethical frameworks, administrative considerations, and the need to safeguard the legitimacy of the tax system?

Based on existing evidence, it is hard to assess whether an EITC is an appropriate policy measure to boost labour supply. Theory predicts that labour supply can be affected through three channels: (a) extensive margin effect—by lowering participation tax rates, the EITC makes taking a job more attractive; (b) income effect—higher disposable income may reduce the hours worked by those who already have jobs; (c) substitution effect—changes in marginal tax rates change the payoff from an extra hour of work. If the EITC is to target low and middle incomes, it needs to be phased out for higher earners. This increases the effective marginal tax rate in the phase-out segment, and both income and substitution effects are expected to reduce labour supply.

EITCs can be designed in a number of ways, varying in size, phase-in, and phase-out. To quantify the expected effects of different designs, the research group started by developing a quantitative labour supply model to predict labour supply responses, both on the extensive and intensive margins, to changes in the tax schedule. The model was applied to microdata for the full population of taxpayers in the relevant age range and combines register-based income and benefit data with parameterised behavioural responses (participation and intensive-margin elasticities), explicit treatment of means-tested benefit tapering, and stochastic participation decisions. Potential earnings for non-participants are imputed as 80% of observed employed peers with the same sex, age, and education.

In the process, the behavioural and fiscal consequences of a number of designs were simulated, before one was recommended. It should come as no surprise that researchers designing an EITC must address certain unavoidable properties:

1. The larger the deduction, the larger the effect on labour market participation.
2. The larger the deduction, the larger the reduction in tax revenue.
3. An EITC that targets the low- and middle-income segment will involve a phase-out. This lessens the reduction in tax revenue but has a negative labour supply effect on the intensive margin.

In order to predict the quantitative importance of these factors, researchers need both a model and data for the income distribution. Needless to say, such predictions involve a high degree of uncertainty.

Table 1 shows the key behavioural parameters that were set exogenously: In the baseline scenario, the participation elasticity was set to 0.5, the compensated elasticity to 0.1, and the income effect to -0.05 .

Table 1. Key behavioural parameters used in the simulations

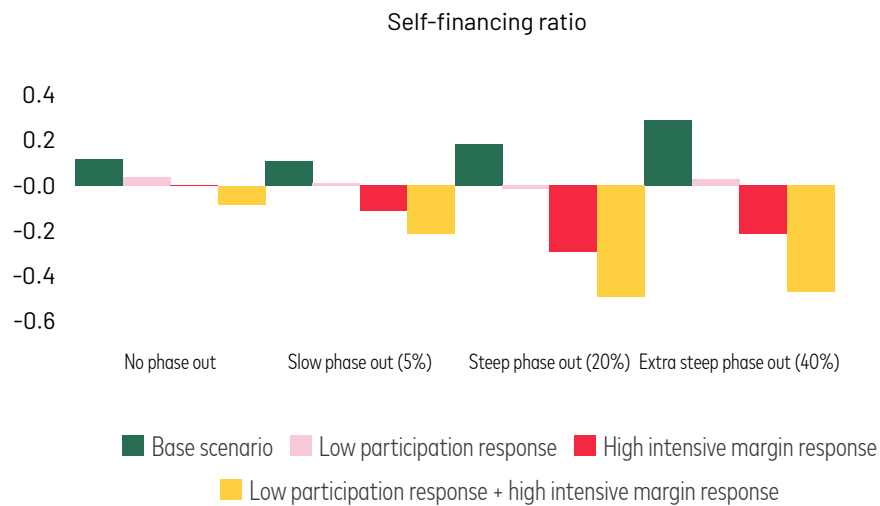
	Participation elasticity	Income effect	Compensated elasticity
Baseline scenario	0.5	-0.05	0.1
Low participation response	0.2	-0.05	0.1
Strong intensive margin response	0.5	-0.2	0.3
Low participation response and strong intensive margin response	0.2	-0.2	0.3

Source: Bensnes et al. (2025)

For a given budget, the government faces a choice: A large deduction combined with a phase-out for higher incomes, or a small deduction without a phase-out. The former should have stronger positive effects on the extensive margin and a stronger negative effect on the intensive margin. Hence, this can be seen as a high-risk strategy, since both the potential gains and costs may be greater than in the alternative that does not involve a phase-out.

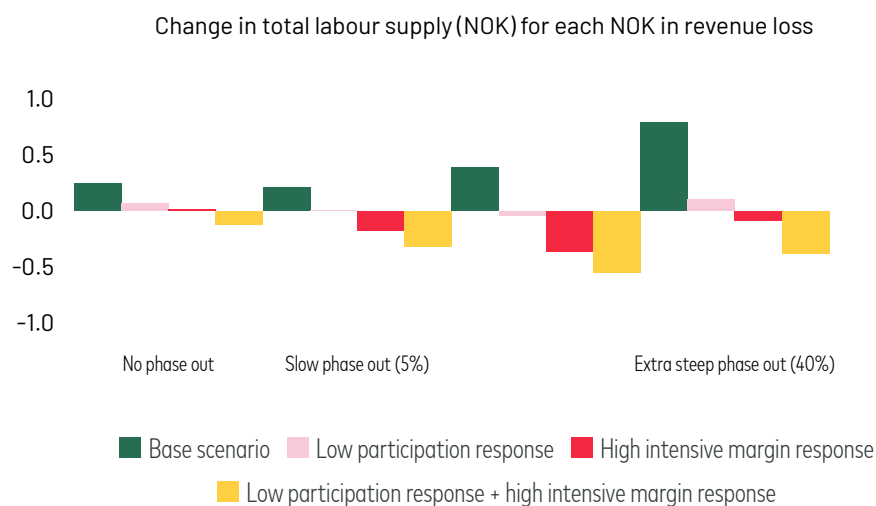
This is also illustrated in Figures 1 and 2, which display the self-financing ratio and the change in labour supply relative to the loss in tax revenue under different phase-out alternatives (no phase-out, slow, steep, and extra steep), and different alternatives regarding the strength of the behavioural parameters in the model.

Figure 1. Estimated self-financing ratios with different designs and under different assumptions regarding labour supply responses



Source: Bensnes et al. (2025)

Figure 2. Estimated change in total labour supply with different designs and under different assumptions regarding labour supply responses



Source: Bensnes et al. (2025)

For three main reasons, the alternative with the steepest phase-out was the one recommended by the research group. First, under the baseline modelling scenario, its expected implications were preferred as a (unspecified) weighted combination of outcomes, including expected changes in labour supply, reductions in net taxes and labour market participation, as well as income inequality. Second, the "risk" of potentially weaker responses on the extensive margin and stronger (negative) responses on the intensive margin was accorded little weight, since the experiment only covers a fraction of the population for a limited time. Third, the "riskiness" in terms of stronger expected responses on both margins is actually a strength, given that the purpose is to gain knowledge about these behavioural changes, as larger responses are easier to separate from "noise" in the data.

Table 2 below summarises the predicted effects from the main parameterisation of the model, using the baseline parameters (cf. Table 1).

Table 2. Simulated behavioural and fiscal consequences of an EITC (extra steep phase-out alternative) for all taxpayers aged 20–35 using baseline behavioural parameters

Expected change in...	
Participation rate	0.8p.p.
No. of people participating in the labour market	9,364 (out of 1.1 million)
No. of people shifting from a minor to a major job (the "super extensive" margin)	5,691 (out of 1.1 million)
Average labour earnings	NOK 1,776
...conditional on earnings < NOK 300k	NOK 10,553
...conditional on earnings > NOK 300k	NOK -4,478
...among disability insurance recipients	NOK 7,841
...among students	NOK 4,305
...social assistant recipients	NOK 22,442
Inequality (Gini after tax)	-0.0098
Fiscal effect	NOK -5,169 mill.
...excluding behavioural responses	NOK -6,015 mill.

Source: Bensnes et al. (2025)

Needless to say, the conclusions from such a simulation exercise depend on both the data and the model parameterisation. We do not know the income distribution in 2026 and later, and we have used 2022 as a proxy. We believe the uncertainty generated by this aspect will be minor. More important is the assumed strength of the responses, the labour supply elasticities, as illustrated in Figures 1 and 2.

The most important lesson from this is that, if the motivation for introducing the tax credit is to increase labour supply, EITCs can easily be counterproductive, i.e., based on existing knowledge, we cannot rule out them reducing the labour supply instead of increasing it.

The simulation model also allows us to describe the expected labour supply responses in detail. It is instructive to divide the population into six different groups, based on how they are expected to change their behaviour because of the tax allowance:

1. Non-participants: Approximately 10% of the population in this age range has no labour income. The EITC reduces their participation tax rate by increasing the net income gain from participating in the labour market. The effect is unambiguously to increase labour supply. With the baseline parameterisation, we predict the tax allowance will increase the participation rate by 0.8 percentage points.
2. "Super-extensive" margin movements among very low earners: A proportion of the population participates in the labour market, but to a very low degree. As such, they are unaffected by the EITC on the intensive margin (their income is below the basic deductions in the tax system). When offered the tax allowance, we predict that some of these individuals will increase their labour supply on the extensive margin in the same way as non-participants.
3. Low earners: For workers with incomes between the basic deductions and the tax allowance for earned income (200k – 320k), the marginal tax rate will decrease substantially, by 22 percentage points. This indicates substantial positive labour supply effects. This group will also have a lower average tax rate and a corresponding reduction in labour supply due to the income effect. In theory, the net labour supply response is thus ambiguous. In our parameterisation, however, the model predicts labour supply will increase.
4. Maximum gainers: For people earning between 320,000 and 335,000, the EITC is maximised, and the marginal tax is unaffected. For this relatively small group, there is no substitution effect, only an income effect. They will gain NOK 27,500, and the model thus predicts a small reduction in labour supply.
5. Medium earners: From 335,000 to 647,000, the allowance is phased out. This is done quite rapidly, such that they will lose NOK 0.4 in allowance per extra 1 NOK in earnings. This 0.4 NOK is taxed at 22%, such that their effective marginal tax will increase by 8.8 percentage points. Hence, for these workers, the theoretical predictions are an unambiguously decreased labour supply, since they have both negative income and substitution effects.
6. High earners: Many in this group earn more than NOK 647,000. These workers are unaffected on the intensive margin. In theory, they could, however, be affected on the extensive margin if, without the tax allowance, they are on the margin between full-time earnings, e.g., 800,000 and part-time earnings, e.g., 400,000. In the simulation model, we have ignored this effect, but an evaluation will show whether this margin is relevant or not.

3.1 Choice of experimental design, and how it relates to legal requirements

Randomised allocation of different tax schedules is at odds with general legal principles of equal treatment. Norway does have legislation specifically for pilot schemes and experiments, but it is mainly for situations in which a municipality or other public-sector body wants to try out new ways of organising parts of their services.

From a legal perspective, the starting point is the Constitution, which states that everybody is equal in the eyes of the law. The key questions are whether differential treatment is objectively justified and proportionate. In this case, the reason for it is the "need for knowledge". A better understanding of behavioural responses to such a tax change is considered important. The latter requirement (proportionality) implied that we needed to investigate whether the same knowledge could be obtained in a less intrusive way, without randomisation or with a lesser degree of unequal treatment.

To address this, two different potential research designs and evaluation strategies were considered that did not involve randomisation:

1. Providing the treatment to everyone aged 20–35 in rural municipalities with low employment rates: As the expected effects are relatively small at the individual level, these effects will be dwarfed by the lack of parallel trends and expected fluctuations when comparing these municipalities with other ones.
2. Providing the treatment to everyone in a narrow age range, e.g., age 25–30: This avoids randomisation but provides opportunities to use quasi-experimental methods. However, this type of evaluation requires additional assumptions, such as common time- or age-effects across cohorts/years. To assess this approach, the research group constructed a panel dataset with annual earnings for a number of years and birth cohorts (77 combinations). Within this dataset, we estimated a number of potential placebo models, assuming treatment was given at different years and different age thresholds. For each combination, we estimated a simple DiD model and saved the coefficients. The standard deviation of these coefficients informs us about what we have chosen to term "design uncertainty", which is a form of uncertainty not captured by the standard error (which captures sampling uncertainty). The design uncertainty arises from the fact that any deviation from the additional assumptions in the DiD model will lead directly to bias in the estimated treatment. It turns out that, in this case, the design uncertainty is about the same size as the expected effects, so we concluded that this would not be a satisfying strategy for achieving the desired knowledge. Note that this assessment is not one of principle, but pragmatic, and related to the case in hand.

The general conclusion from this assessment was that we could only fulfil the knowledge needs if the research design was based on some kind of randomisation. This was again crucial for the legal assessment conducted by the Ministry of Finance, which concluded that the requirements for unequal treatment were satisfied.

A randomised experiment can be designed in many ways. In the first assessment, two alternatives were considered: Randomisation of municipalities and randomisation of individuals. The first has some advantages as it also captures equilibrium effects, at least in theory. However, this is only true if the unit of randomisation is aligned with the entities in which the general equilibrium is reached, e.g., independent labour markets. Norway has 358 municipalities, many of which are very small. The number of local labour markets depends on the definition used, but commonly used categorisations vary between 40 and 130 local labour markets.

To compare the two alternatives involving randomisation, we conducted power calculations for stratified randomisation. The strata are based on year of birth, gender and previous earnings, which substantially increases precision. We find that with individual randomisation, the minimum detectable effect for labour market participation is 0.4 percentage points. By contrast, even if the experiment involves 50% of all municipalities (approx. 550,000 treated), the minimum detectable effect in a municipality-based randomisation design is no less than 0.65 percentage points. Hence, in order to obtain the necessary precision using this approach, the study would have to involve a large number of treated subjects and increase the fiscal costs by an order of magnitude. The combination of statistical precision and fiscal cost made the research group recommend the alternative using individual randomisation.

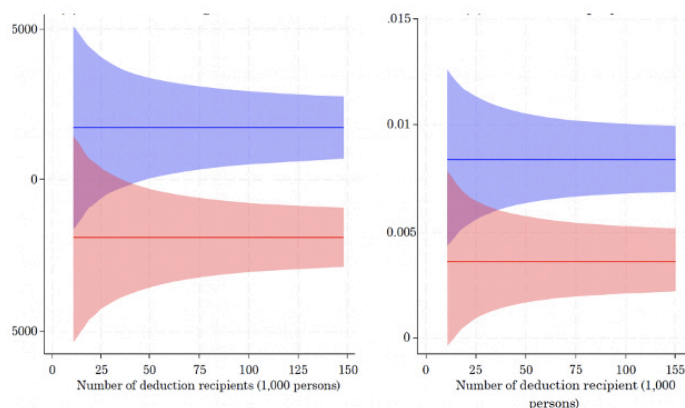
At a later stage, the research group also considered a mix of approaches, with varying treatment intensities between municipalities. This is a design used by, among others, Crépon et al. (2013) and Cheung et al. (2024), and which, in many respects, captures the best of both worlds with individual randomization and an opportunity to learn about spillovers and equilibrium effects. However, the research group decided in the end that the potential for useful insights about such effects within a realistic budget was too limited to recommend such a design.

The research group also had to provide a recommendation about the size of the treatment group, as well as the number of treatment arms. As mentioned above, the minimum detectable effect for the participation rate is actually quite small, even with, e.g., 35,000 in the treatment group. However, to assess responses along the intensive margin and participation effects in various subgroups, the research group recommended a treatment group of 100,000.

Figure 3 illustrates the expected effects on a) average labour income (left) and b) employment (right) according to various assumptions about the magnitude of the

responses, as discussed above. The shaded areas (95% confidence intervals) show how precision increases with the number of recipients. With 100,000 people receiving the deduction, it is expected that the responses can be estimated quite precisely.

Figure 3. Expected effects on average labour income (left) and employment (right) according to different assumptions regarding labour supply responses (cf. Table 1) and 95% confidence intervals



Behind this recommendation also lies a degree of caution. Many things can go wrong, and the researchers wanted to avoid an outcome with a number of underpowered experiments offering limited learning opportunities. On the number of treatment arms, the research group recommended three – one with a different-sized tax credit, and another with a different phase-out.

The final recommendations from the research group can be summarised as follows. The expected effects of introducing an EITC are highly uncertain. Unless the EITC is introduced experimentally, the effects cannot be credibly estimated. In order to gain sufficiently precise estimates, the experiment should be based on individual randomisation and be large, with around 100,000 recipients per treatment arm. The experiment should ideally include several treatment arms, e.g., with and without a phase-out. However, the research group also presented a preferred alternative for one treatment arm.

4 Public hearing and debate

The proposed experiment required specific legislative changes and a public hearing. It was also clear that the experiment would be extraordinary in scale, and that the use of randomisation and experimental research methods deserved to be the subject of public debate.

The experiment was first presented at a press conference on 27 May 2025 by the Minister of Finance, Jens Stoltenberg, the Minister of Labour and Social Inclusion, Tonje Brenna, and the research group. Our impression is that the experiment came as a surprise to the media, and not all news outlets even bothered to turn up. This changed rapidly, and what has been termed “Stoltenberg’s Tax Lottery” became a big news story. It quickly also became controversial and was criticised – and praised – by key media commentators and several political parties. Initially, the discussion went along two lines: First, that the use of randomisation in combination with taxes was unfair and could undermine trust in the tax authorities and system. Second, that it was unnecessary, either because critics expected the EITC not to work or were certain it would.

During the public hearing, which was due on 1 August, the Ministry of Finance received a large number of consultative comments. A summary can be found in the Proposal to the Parliament (in Norwegian) here: [Prop. 1 LS \(2025–2026\) - regjeringen.no](https://www.regjeringen.no).

5 From idea to reality: The practical implementation and the research design used

The budget for 2026 was not passed in the Parliament until December 2025. Hence, in order to start the experiment on 1 January 2026, the practical implementation needed to be undertaken without knowing whether the experiment would be approved. Below, we will present the different elements of the experiment, which is now up and running.

5.1 The tax scheme

The EITC in the experiment is similar to the research group’s main alternative. Compared to previously proposed tax deductions of this type in Norway, it is relatively sizeable, with a maximum value of NOK 125,000. It raises the threshold for paying the 22% tax on ordinary income from around NOK 210,000 to 335,000.

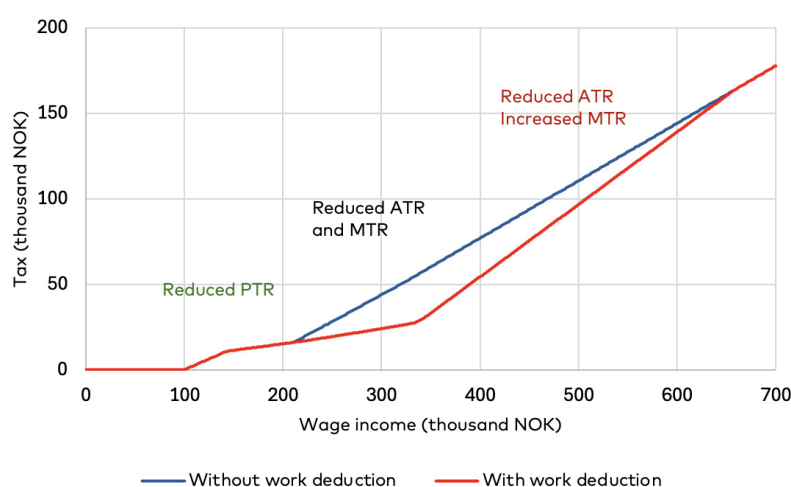
It also targets low and middle incomes. For every NOK earned exceeding NOK 345,000, the allowance is reduced by NOK 0.40. This increases the effective marginal tax rate (MTR) by 8.8 percentage points in the phase-out region. The ordinary scheme and the test scheme with this deduction are illustrated in Figure 4.

The EITC is fully phased out when labour income exceeds NOK 657,500, which is somewhat below the average income in Norway. Within the target group for the experiment (20–35 years), around 31% of taxpayers are expected to have income above this threshold. For another 28%, earned income is zero or net income is expected to be below the threshold for paying the 22% “tax on ordinary income”. Hence, this group will

not benefit from the allowance unless they earn more. However, their participation tax rate (PTR) will be lowered, increasing their incentives to start work or to increase their labour supply.

For the remaining 42% of the target group, the allowance will reduce their average tax rate (ATR). For those with earnings below NOK 335,000, the marginal tax rate will also be reduced, while for those with earnings exceeding NOK 345,000 in the phase-out region, it will increase their marginal tax rate. See Figure 4.

Figure 4. The tax schedule with and without the EITC. Wage earner with standard deductions and only wage income. 2026-rules



5.2 The "lottery" draw

The random distribution of deductions was carried out by computer code. This code was made public along with the budget proposal. For practical reasons, related to the survey (see below) and to the planning of individual tax cards, the random draw into treatment and control groups needed to be carried out before Parliament passed the budget. We ended up with this process:

- A. The computer code was made public on the Ministry of Finance website ([here](#)). The code specified the draw conditional on two "keys". The first key was a code constructed as a function of the dataset ("hash"), which ensures that the dataset used is unchanged. The second key was the (at the time, not yet realised) opening value of the Oslo Stock Exchange (OSEBX) on 3 November 2025.

- B. The dataset used for the draw was constructed by the tax administration on 31 October, before the second key was available. This guarantees that those constructing the dataset in the tax administration were unable to manipulate the outcome of the draw by "fiddling" with the data.
- C. On 3 November, the opening value of OSEBX was added as the second key and the draw was made.

The relevant population contains approximately 1.3 million taxpayers, and 8% were randomised into the treatment group. The randomisation was "stratified," with the population divided into small and homogeneous groups before the draw. The groups were constructed by first grouping the population based on:

- Year of birth
- Gender
- Whether or not they received social security benefits in 2025 (January to September)
- Whether or not they had labour earnings in 2024
- Whether or not they had self-employment/business income in 2024
- Residency status and tax liability status in Norway.

Within each of these groups, the population was sorted by their labour earnings in 2025 (January to September) and divided into groups of 100. From each stratum, eight individuals were randomly drawn to be in the part of the tax system with the deduction. This group is denoted group A. The remaining 92 individuals were randomly split into two equally sized groups, groups B and C, one that will receive an information treatment (B) and one that will not (C).

5.3 The information given

Everyone involved received information about the experiment at the same time as the annual tax cards were made available. The tax authorities sent text messages or emails saying that "Your tax card for 2026 is ready. This year, the government is starting a trial with work deductions for young people. Log in to our website to see if you are eligible for a work deduction."

After logging in, a "banner" appears about the experiment. When you click on it, you receive the information treatment described below. The system registers whether or not you visit the information page, and this information is used to separate "compliers" from "non-compliers" in the evaluation.

Behavioural responses to a tax change require that people also change their perception or belief of what their taxes are. Following a regular tax reform, we believe such changes take time. However, the experiment has a limited duration, and we do not have time to let the new tax system influence perceptions in a natural way. Also, since the reformed tax system only applies to a minority of taxpayers, the change in perceptions will be different from changes in a normal situation.

When designing the experiment, we aimed to simulate changes to perceptions in the long term following a reform. However, we need to achieve this almost instantly, or at least within a short time frame.

To obtain this, the treatment group is provided with the following five pieces of information (see also Figure 5):

1. Predicted income for the next year (used to calculate each person's tax schedule for advance tax).
2. Predicted tax, including the expected tax reduction from the EITC.
3. Individual specific examples on the amount of additional tax – and resulting increase in net income – stemming from:
 - a. an increase in labour earnings of NOK 1,000. We think of this as informing recipients of the marginal tax rate without using the term.
 - b. an increase in labour earnings of NOK 100,000. We think of this as the "participation tax rate" for a medium-sized income change.
 - c. an increase in labour earnings of NOK 300,000. We think of this as the "participation tax rate" for individuals outside the labour force.

The extraordinary information given to the treatment group presents us with a second challenge: That the information regarding marginal and participation tax rates could in itself affect labour supply.

To disentangle the effect of information from the effect of the deduction, we split the control group randomly into two parts. Half of them (Group B) receive the same information but with numbers calculated for the ordinary tax scheme, see Figure 6. The others (Group C) received just a minimum of information, see Figure 7.

Figure 5. Information given to group A

Work allowance for young people

About the pilot scheme

From 2026, the government is launching a [pilot scheme for a work allowance for young people \(opens in new tab\)](#). Everyone born between 1991 and 2006 is included in the target group, and around 100,000 people have been selected to receive the allowance for the duration of the pilot.

 You've been randomly selected and are entitled to the work allowance

If you increase your employment income in 2026

How much tax you pay depends on how much you earn.

We've pulled some figures from your tax deduction card so you can see how much tax you'll pay if you work more – and you're entitled to the work allowance.

The allowance is calculated automatically and may reduce the tax you pay on employment income.

Key figures from your tax deduction card

We've estimated in your tax card that you'll earn **NOK X** in 2026. The work allowance reduces your tax by **NOK X**. Your total tax payable is **NOK X**.

[^ What the key figures mean](#)

- **Income** includes all salary, benefits, capital income, rental income and more.
- **Tax** is the total amount we've calculated based on the income and deductions listed in your tax deduction card.
- **Reduced tax** is what you save due to the work allowance.

Check that your income and other details in your tax card are correct – so you do not pay too much or too little tax.

[→ View or update your tax deduction card](#)

If you increase your employment income by:

NOK 1 000

You'll pay **NOK X** more in tax and keep **NOK X** of the increased income.

NOK 100 000

You'll pay **NOK X** more in tax and keep **NOK X** of the increased income.

NOK 300 000

You'll pay **NOK X** more in tax and keep **NOK X** of the increased income.



Figures are based on your tax deduction card. Check that your income and other details are correct – so you do not pay too much or too little tax.

[→ View or update your tax deduction card](#)

Figure 6. Information given to group B

Work allowance for young people

About the pilot scheme

From 2026, the government is launching a [pilot scheme for a work allowance for young people \(opens in new tab\)](#). Everyone born between 1991 and 2006 is included in the target group, and around 100,000 people have been selected to receive the allowance for the duration of the pilot.



You have not been randomly selected and are not entitled to the work allowance

If you increase your employment income in 2026

How much tax you pay depends on how much you earn.

We've pulled some figures from your tax deduction card so you can see how much tax you'll pay if you work more.

Key figures from your tax deduction card

We've estimated in your tax card that you'll earn **NOK X** in 2026. Your total tax payable is **NOK X**.

[^ What the key figures mean](#)

- **Income** includes all salary, benefits, capital income, rental income and more.
- **Tax** is the total amount we've calculated based on the income and deductions listed in your tax deduction card.

Check that your income and other details in your tax card are correct – so you do not pay too much or too little tax.

[→ View or update your tax deduction card](#)



If you increase your employment income by

NOK 1 000

You'll pay **NOK X** more in tax and keep **NOK X** of the increased income.

NOK 100 000

You'll pay **NOK X** more in tax and keep **NOK X** of the increased income.

NOK 300 000

You'll pay **NOK X** more in tax and keep **NOK X** of the increased income.



Figures are based on your tax deduction card. Check that your income and other details are correct – so you do not pay too much or too little tax.

[→ View or update your tax deduction card](#)

Figure 7. Information given to group C

Work allowance for young people

About the pilot scheme

From 2026, the government is launching a [pilot scheme for a work allowance for young people \(opens in new tab\)](#). Everyone born between 1991 and 2006 is included in the target group, and around 100,000 people have been selected to receive the allowance for the duration of the pilot.

i You have not been randomly selected and are not entitled to the work allowance

5.4 The survey

In parallel with the experiment, we are also conducting a large survey in two waves. The first wave was in November 2025, when 30,000 participants were invited to take part. The second wave will be in 2026 and include all respondents from the first wave. In addition, 50,000 new participants will be invited to take part. In the evaluation, the survey data will be linked to administrative data.

Among other things, the survey is designed to inform us about respondents' perceptions of their tax rate, both before and after the experiment has started. Perhaps most importantly, it will inform us on whether their perceptions of their tax rates differ, after treatment, between groups A, B and C.

The participants are also asked questions about trust in the tax authorities and others, facilitating at least some investigation of some of the critical concerns raised in the public debate and hearing.

5.5 Consent and research ethics

Both the Ministry of Finance and the research group have a responsibility to assess ethical concerns arising from the research and abide by established research ethical norms. Several such concerns have been raised, both by the research group and the public. We will limit the discussion here to the arguably three most important concerns^[3]:

3. The research groups full ethical assessment is available online: https://frischsenteret.no/publication-details/skrift_id=1790

1. Negative effects on those who gain nothing: Research is not supposed to harm anybody. This experiment will do so indirectly by providing extra resources to some at the expense of others. It is our assessment that this negative effect will be so small as to be negligible. In principle, this is the case for all experiments paid for out of the public purse.
2. Harmful disappointment: Will the control group be disappointed, and might this have a negative effect on them? Again, we believe the effect will be negligible. We also believe the disappointment of losing in a lottery with a small probability of winning is much less than in one with a high probability of winning, say 75%.
3. Active or passive consent: The main rule in the Norwegian ethical guidelines for social science research is that participants should provide active consent.^[4] In theory, everyone in the target group could be asked to read an online invitation and actively sign up (digitally) to confirm they want to take part. It is our assessment that such a requirement would substantially reduce the value of the research. Those who consent to take part might differ substantially from the full population, and the results of an experiment involving this group might not be generalisable.

Another option is passive consent, where individuals can reserve themselves from taking part in the experiment. This is the approach preferred by the research group. However, the Ministry of Finance concluded that asking the full target group to take part (or not) in a project not yet approved by Parliament would be disrespectful to the legislature. Given this, the research group and the Ministry of Finance agree that the experiment entails no negative risk to anyone (participants can only gain), and that it is acceptable to conduct the experiment without consent.

6 The way forward

The experiment with an earned income tax credit for young adults represents a major investment in evidence-based labour market policy that aims to fill critical knowledge gaps about how tax incentives influence labour supply in a Nordic welfare state. The tax experiment has just started, and no results are available yet, neither to the researchers nor the public. We believe this experiment will provide important insights and hopefully also expand the possibility frontier for future experiments.

4. [Guidelines for Research Ethics in the Social Sciences and the Humanities | Forskningsetikk](#)

References

- Bensnes, S., et al. (2025). *Utredning om forsøksordning med arbeidsfradrag*. Retrieved from: <https://.../Utredning-om-forsøksordning-med-arbeidsfradrag.pdf>
- Cheung, M., Egebark, J., Forslund, A., Laun, L., Rödin, M., & Vikström, J. (2024). Does job search assistance reduce unemployment? Evidence on displacement effects and mechanisms. *Journal of Labor Economics*, 43(1), 47–81.
- Crépon, B., Duflo, E., Gurgand, M., Rathelot, R., & Zamora, P. (2013). Do labor market policies have displacement effects? Evidence from a clustered randomized experiment. *The Quarterly Journal of Economics*, 128(2), 531–580. Retrieved from: <https://doi.org/10.1093/qje/qjt001>
- Gopinathan, U., Elgersma, J., Dalsbø, T., Bjørbæk, M., & Fretheim, A. (2025). Strengthening research preparedness for crises: Lessons from Norwegian government agencies in using randomized trials and quasi-experimental methods to evaluate public policy interventions. *Health Research Policy and Systems*, 23(1), 8.
- Graber, M., Mogstad, M., Torsvik, G., & Vestad, O. L. (2022). Behavioural responses to income taxation in Norway. *Memorandum No. 04/2022*.
- Kleven, H. (2024). The EITC and the extensive margin: A reappraisal. *Journal of Public Economics*, 236, 105135.
- Kostøl, A. R., & Mogstad, M. (2014). How financial incentives induce disability insurance recipients to return to work. *American Economic Review*, 104(2), 624–655.
- Laun, L. (2019). *Ett effektivare jobbskatteavdrag?* Rapport till Finanspolitiska rådet 2019/2.
- Meyer, B. D., & Rosenbaum, D. T. (2001). Welfare, the earned income tax credit, and the labor supply of single mothers. *The Quarterly Journal of Economics*, 116(3), 1063–1114.
- Michalopoulos, C., Robins, P. K., & Card, D. (2005). When financial work incentives pay for themselves: Evidence from a randomized social experiment for welfare recipients. *Journal of Public Economics*, 89(1), 5–29.
- Verho, J., Hämmäläinen, K., & Kanninen, O. (2022). Removing welfare traps: Employment responses in the Finnish basic income experiment. *American Economic Journal: Economic Policy*, 14(1), 501–522.

Comments on Simen Markussen and Marie Bjørneby: The Norwegian Tax Experiment

Kaisa Kotakorpi

1 Introduction

Markussen and Bjørneby (2026) present the design of a tax experiment to be carried out in Norway. The experiment involves a randomised natural field experiment, or RCT, in which 8% of 20–35-year-old Norwegian taxpayers are eligible for an earned income tax allowance, while all other taxpayers see no change to their tax schedule. The experiment affects the participants' marginal tax rates and participation tax rates. I will not review the details of the experiment here, but refer to the account provided in Markussen and Bjørneby (2026). The purpose of this comment on the experiment is twofold. First, to look at the Norwegian tax experiment in the context of the development of research involving randomised field experiments in public economics. Second, to comment on the attractive features of the experiment and raise a few questions about the design that deserve consideration when the findings are discussed and interpreted in due course.

2 Context: field experiments in public economics

Natural field experiments or randomised controlled trials (RCTs) conducted in real-world settings have proliferated as part of the credibility revolution in empirical economics. Public economics and, in particular, research on taxation have lagged behind somewhat. RCTs appear to have been more common, e.g., in development, education, and labour economics. This methodological development has been documented in the context of public economics by Kleven (2014). I will not attempt a comprehensive review of the literature or a treatment of the multitude of issues that have to be considered when conducting this type of research. Rather, I will make a few remarks that put the Norwegian tax experiment into context.

One prominent reason why field experiments have not been commonly used to analyse the effects of taxation is probably that randomising tax rates in the real-world has usually been considered unethical or inequitable from the point of view of horizontal equity, i.e., against the principle of equal treatment. However, this argument should not lead to an outright ban on tax experiments. A comprehensive discussion is beyond the scope of this comment. One point worth considering is that any policy change that does not apply to the entire population will require some lines to be drawn, and this will

entail at least quasi-randomness in the treatment of similar individuals, depending on which side of the eligibility line they happen to fall. In addition, while compromising some equity is a cost, experimentation also has benefits, e.g., generating reliable knowledge about the effects of major policy programmes and hence avoiding misguided choices that may prove costly. Therefore a cost-benefit calculus on the use of experimentation may be warranted. (Of course, there may also be principles that should not be compromised at any cost – a type of argument that requires clear justification.)

However, the question of the ethics of randomisation does not mean that no field experiments have been conducted in the fields of public economics or tax research. One particular strand of research in which they have been prominent and useful is the study of tax evasion and enforcement. Conducting a randomised experiment on these topics has probably been viewed as less problematic from an ethical point of view, as randomised (tougher) enforcement should affect only those who have not complied with their legal duties as taxpayers. One benefit of using field experiments to analyse tax evasion is that innovative research designs are particularly valuable in uncovering evasion. It is something that, by definition, the individuals concerned want to hide, so it is not immediately observable in administrative data. However, reactions to randomised enforcement measures can provide information about tax evasion in the baseline. Disadvantages of this approach include the need for careful interpretation of the findings, as real-life enforcement, by its very nature, targets a heavily selected group of taxpayers. As such, lessons about the effects of randomised tax audits, for instance, are unlikely to be directly generalisable to tax administrations' real-life risk-based audit programmes.

A review of this strand of literature is provided by Pomeranz and Vila-Belda (2019), and a meta-analysis by Antinyan and Asatryan (2024). In the Finnish context, papers that report results from randomised field experiments related to tax enforcement include Kosonen and Ropponen (2015) on providing information about a VAT reform; Harju et al. (2020) on evasion of taxes on imported used cars; and Eerola et al. (2025) on evasion of rental income tax. Insights into cooperation between researchers and tax administrators in Finland and the UK are discussed in Almunia et al. (2019).

The papers mentioned above applied randomised variation to other design features of the tax system, but not tax rates. Experiments explicitly randomising tax rates are rare. A prominent Finnish example that involves randomisation of effective tax rates is the Finnish Basic Income Experiment. The design and effects of the experiment have been reported in a series of papers, Verho et al. (2022), Hämäläinen and Verho (2022) and Hämäläinen et al. (2025). In applying randomised eligibility to basic income for a subset of unemployed Finns, the experiment created randomised variation in participation tax rates, i.e., the fraction of earnings individuals get to when they move from unemployment to work.^[5] This provides a unique setting in

5. References to relevant earlier research are cited in these papers and include studies on the negative income tax experiments in the U.S. in the 1970s and randomised in-work benefits in the Self-Sufficiency Project (SSP) in Canada in the 1990s.

which to examine the causal effects of work incentives on outcomes such as labour force participation.

Implementing the Finnish basic income experiment was a major achievement from both scientific and societal perspectives, and it has profound implications for what is feasible for future generations of researchers and public administrators. The design of the experiment and the ethical considerations behind this type of experimentation were reviewed and approved by the Constitutional Committee of the Finnish Parliament. This showcases the importance of the experiment and the process, as the Committee assessed whether this type of randomised experiment is consistent with citizens' constitutional rights. Indeed, the experiment has paved the way to other important endeavours that have followed, such as Finland's Two-Year Preschool Experiment, also reviewed in this volume (Izadi and Sarvimäki, 2026).

3 Comments: design of the Norwegian tax experiment

The Norwegian tax experiment is important in the context of the background described in Section 2. It applies randomised variation to the earned income tax allowance to provide credible causal evidence on the effects of an important feature of income tax systems in many countries. Discussions during the planning stage of the experiment involved careful consideration of ethical issues and the benefits and drawbacks of experimentation, as discussed above. One feature of the experiment I will comment on below is that it combines variation in effective tax rates with a personalised information treatment, making the allowance's effect on effective tax rates salient to the participants.

The experiment has multiple attractive features. First, as should be clear by now, this type of research is very rare, and an RCT with randomised variation in effective tax rates is exceptional. Second, a key feature is that the experiment induces variation in both marginal tax rates and participation tax rates, which allows for the study of both extensive and intensive margin effects on labour supply. Third, the experiment is on a very large scale, involving 8% of taxpayers in a wide age bracket. Fourth, the experiment lasts for a relatively long time, i.e., five years, allowing time for the effects to emerge if participants take a while to react, e.g., if new jobs cannot be found immediately when the incentives change. Fifth, there are no other restrictions on the target group beyond the age restriction, providing good generalisability of the results to a broad, well-defined subgroup of the population. Sixth, the researchers intend to measure perceptions of tax rates before and after the treatment. This is very important for interpreting the results of the experiment: any effects will be conditional on successfully changing individuals' perceptions of tax rates. In other words, the strength or impact of the treatment depends on changes in perceptions; this is the first stage variation that induces (or does not) changes in the final outcomes of

interest. Finally, the experiment provides an opportunity to study the effects on multiple other outcomes, not just direct labour market outcomes.

Are there potential caveats or complications that should be kept in mind when discussing and interpreting the results of the experiment in due course? First of all, it is worth considering whether introducing a tax allowance rather than changing nominal tax rates directly is an innocuous choice. Referring to a “tax-cut experiment” perhaps carries with it the connotation of cutting nominal tax rates, whereas in practice the experiment involves introducing an allowance. This, of course, implies a cut in effective tax rates, but given that a tax allowance are highly complex, the salience of cuts implemented to tax rates directly vs. indirectly via an allowance may be different. On the other hand, combining the tax allowance with explicit information about participants’ actual tax rates will make the tax changes implied by the experiment salient. This makes cuts implemented via an allowance look more like a cut to nominal tax rates.

A second and related question concerns the external validity of this type of combined treatment, i.e., a tax cut/allowance combined with highly personalised information. In real life, policy changes are typically not combined with the provision of such extensive and personalised information directly to individual taxpayers. In this sense, non-compliers in the combined treatment who receive both the tax cut and the information but do not access the information are of interest, as they are closest to a real-life situation in which policy changes are implemented without a simultaneous change in information. This, however, is a highly selected group.

The plans for the experiment do not include a separate treatment with *only* a tax change and no information. In my view, this type of pure tax treatment would have been of interest. The argument appears to be that having a combined treatment with information and a tax change, as well as a separate treatment with information only, would allow for the separation of the effects of the two. However, as all participants in this constellation do receive information that makes tax rates salient, this does not necessarily allow researchers to pin down the precise effects of tax changes in the *absence* of such information. Rather, all effects of the experiment are measured *conditional* on receiving the information treatment; whether they would be the same in the absence of information is not clear a priori, or at least it warrants discussion. One intriguing line of thought in this context is whether the problem lies with the experiment or with the real world – why not provide better information in the real world as well? Could tax rate changes routinely be accompanied by personalised information to increase their salience to taxpayers? Why this is not done in practice may, of course, be due to multiple reasons, such as implementation costs, potential problems with information accuracy related to complex policy changes, and, more fundamentally, it is not even obvious whether full salience is always optimal (see Goldin (2015) for a discussion on optimal tax salience).

Third, the control group in the experiment is not fully “business-as-usual”: the individuals in the control group receive an invitation to log in and find out whether they were selected for the experiment. This may trigger behavioural reactions to the possibility of being eligible for a tax allowance, only to find out that, in fact, you are not receiving it.

Fourth, an important consideration is the size of the changes in incentives implied by the experiment. For example, the changes in participation tax rates in the Finnish Basic Income experiment were very big. The experiment lowered participation tax rates by 23 percentage points for full-time employment, yet no significant change was found in the number of days in employment (Verho et al. 2022). Even though the sample size in the Norwegian experiment is an order of magnitude larger, meaning greater power to detect even small effects, the magnitude of the treatment may not be solely a power issue; reactions to small vs. large incentive changes may be genuinely different, e.g., due to labour market frictions.

All in all, the Norwegian tax experiment is an important endeavour both scientifically and from a societal and public administration perspective. The results and careful discussion of their interpretation and implications will be of interest to both the scientific audience and policymakers.

References

- Almunia, M., Harju, J., Kotakorpi, K., Tukiainen, J., & Verho, J. (2019). Expanding access to administrative data: The case of tax authorities in Finland and the UK. *International Tax and Public Finance*, 26(3), 661–676.
- Antinyan, A., & Asatryan, Z. (2024). Nudging for tax compliance: A meta-analysis. *The Economic Journal*, 135(668), 1033–1068.
- Eerola, E., Kosonen, T., Kotakorpi, K., & Lyytikäinen, T. (2025). Tax compliance in the rental housing market: Evidence from a field experiment. *American Economic Journal: Economic Policy*. Advance online publication.
- Goldin, J. (2015). Optimal tax salience. *Journal of Public Economics*, 131, 115–123.
- Harju, J., Kosonen, T., & Slemrod, J. (2020). Missing miles: Evasion responses to car taxes. *Journal of Public Economics*, 181, 104096.
- Hämäläinen, K., Simanainen, M., & Verho, J. (2025). Health effects of cash transfers: Evidence from the Finnish basic income experiment. *Journal of Public Economics*, 250, 105480.
- Hämäläinen, K., & Verho, J. (2022). Design and evaluation of the Finnish basic income experiment. *National Tax Journal*, 75(3), 573–596.
- Izadi, R., & Sarvimäki, M. (2026). Learning to experiment: Practical lessons from Finland's two-year preschool experiment. *Nordic Economic Policy Review*.
- Kleven, H. J. (2014). How can Scandinavians tax so much? *Journal of Economic Perspectives*, 28(4), 77–98.
- Kosonen, T., & Ropponen, O. (2015). The role of information in tax compliance: Evidence from a natural field experiment. *Economics Letters*, 129(C), 18–21.
- Pomeranz, D., & Vila-Belda, J. (2019). Taking state-capacity research to the field: Insights from collaborations with tax authorities. *Annual Review of Economics*, 11, 755–781.
- Verho, J., Hämäläinen, K., & Kanninen, O. (2022). Removing welfare traps: Employment responses in the Finnish basic income experiment. *American Economic Journal: Economic Policy*, 14(4), 501–522.

Comments on Simen Markussen and Marie Bjørneby: The Norwegian Tax Experiment

Helena Keinänen, Jarno Tuimala and Olavi Tuomi

Norway is conducting a large-scale, real-world experiment on earned income tax allowances from 2026–2030. The study will involve 100,000 randomly selected taxpayers aged 20–35, who will receive a new deduction. The aim of the randomised, controlled trial (RCT) is to examine how new tax incentives influence labour supply and employment. The Finnish Tax Administration (FTA) regularly uses RCTs to evaluate the efficiency of its activities (Hyytinen et al., 2022). It is well known that RCTs offer an attractive study design characterised by simplicity and clarity. However, if a tax experiment analogous to that in Norway were implemented in Finland, certain questions would need to be considered in advance: The FTA's statutory duties include tax collection, tax control, and safeguarding taxpayers' rights. It must also ensure correct and consistent taxation, develop its service capabilities and perform certain other duties assigned to it by the Ministry of Finance. Academic research is not one of the FTA's responsibilities, and a study similar to the one in Norway could not be implemented. In addition, due to strict constitutional requirements (The Constitution of Finland (731/1999) 81 §), any tax experiment must be based on a specific piece of legislation that stipulates clear rules on tax liability, calculation bases, and the legal remedies available to the persons or entities liable to taxation. They cannot rely on broad administrative discretion or open-ended clauses citing "special reasons".

To conduct a tax experiment equivalent to that in Norway, the political decision-makers and the Finnish Ministry of Finance would need to support the initiative. Implementing the experiment would require new legislation authorising and detailing the trial. In practice, the statute would have to: 1) state compelling public interest and a clear evidence-generation purpose, ensuring that fundamental rights are not infringed; 2) be enacted by Act of Parliament and set precise rules on the target group, selection/area criteria, the scope of the trial, and the transparency of the randomisation method (including code); 3) observe proportionality and necessity, allowing only those deviations from ordinary law that are needed, and avoiding undue advantages or excessive gaps between treatment and control; 4) apply fair, non-discriminatory selection; person-based limits (e.g., by age) are acceptable only when well-justified and non-arbitrary; 5) set out the geographic/target scope and duration in the legislation, with a strict, time-limited mandate; and 6) pre-plan evaluation, governance and reporting to ensure credible, policy-relevant results (Finlex, n.d.-a, n.d.-b; Kangas & Pulkka, 2016; Raitano et al., 2021).

Sufficient funding would be needed as the experiment would entail development needs for the FTA's information and taxation systems. An RCT would also require sufficient staffing of the FTA's customer services, as demand for telephone services and physical visits to its service points would probably increase. The FTA would also need to draft detailed guidance regarding the RCT and train its staff. Long-term technical change management would also be required, given that taxpayers would retain the right to appeal their previous years' tax assessments for several years after the end of the experiment.

However, randomised controlled trials raise a more fundamental issue: the unequal treatment of taxpayers inherent in the RCT design. The Finnish constitution explicitly supports equal treatment for all, a position reinforced by other legislation, such as the Tax Administration Act (Finlex, 2010), which stipulates that the FTA must promote consistent and equitable taxation. As RCTs inevitably result in differential treatment between the experimental and control groups, researchers must address the tension between the need for the high-quality knowledge RCTs provide and the need to protect basic rights. In Finland, the Constitutional Law Committee is the parliamentary body that pronounces on the constitutionality of bills. At the time of the widely publicised basic income experiment in Finland (Helsinki University Library, n.d.), the Committee considered randomised assignment and differential treatment permissible because the enabling Act set out, in sufficient detail, the experiment's objective, target group, selection and randomisation criteria (including the transparency of the method), scope, and duration, along with legal safeguards. Consistent with its broader practice (e.g., in the trial of relief from social security contributions for temporary employers), the Committee accepted randomisation and unequal treatment only when these core parameters were defined by statute with adequate precision and openness. In the basic income case, compulsory participation was also deemed acceptable to secure reliable results, provided the statutory justifications were sufficiently robust (Finlex, 1999).

Before conducting a similar experiment in Finland, it is important to assess whether the study would yield sufficient new knowledge and insights to justify the effort. It is essential to determine whether Finnish and Norwegian societies are sufficiently comparable in the underlying causes of youth unemployment, particularly whether it stems primarily from a lack of willingness to work rather than from structural barriers to employment. If the situations are sufficiently similar, the findings from the Norwegian experiment could be applied directly in Finland without the need for a new trial.

It might be possible to mitigate the inequality inherent in RCTs by adjusting the experimental design. Randomisation could be applied multiple times during the study, increasing the likelihood that more participants benefit from lower tax rates, though possibly for a shorter period. Finally, at the conclusion of the experiment, the control group could also be offered the trial group's tax rate for a fixed period, thereby reducing the disparities caused by randomisation.

Should Finland proceed with a similar experiment, it could plausibly be designed to generate more comprehensive knowledge than the Norwegian trial. For instance, multiple experimental groups could be included to examine taxpayer behaviour under different combinations of tax rates and income levels. The study could also provide insights into how people comply with their tax responsibilities when rates vary.

The experiment Norway is conducting is expected to provide valuable insights into how targeted tax system adjustments may influence employment opportunities for young adults, and its findings will be closely studied in Finland.

References

- Finlex. (1999). *Laki lääketieteellisestä tutkimuksesta 488/1999*. Retrieved from: <https://www.finlex.fi/fi/lainsaadanto/1999/488>
- Finlex. (2010). *Laki Verohallinnosta 503/2010*. Retrieved from: <https://www.finlex.fi/fi/lainsaadanto/2010/503>
- Finlex. (n.d.-a). *Opas kokeilulakien valmistelun tueksi*. Retrieved from: <https://kokeiluohje.finlex.fi/>
- Finlex. (n.d.-b). *Verosta säätäminen (11.3)*. Retrieved from: <https://lainkirjoittaja.finlex.fi/11-verot-ja-maksut/11-3/>
- Hyytinen, A., Tuimala, J., & Hammar, M. (2022). Enhancing the adoption of digital public services: Evidence from a large-scale field experiment. *Government Information Quarterly*, 39(3), 101687. Retrieved from: <https://www.sciencedirect.com/science/article/pii/S0740624X2200020X>
- Kangas, O., & Pulkka, V.-V. (Eds.). (2016). *Ideasta kokeiluihin: Loppuraportti perustulokokeilun toteuttamisvaihtoehdoista*. Valtioneuvoston kanslia. Retrieved from: <https://tietotarjotin.fi/julkaisu/136348/ideasta-kokeiluihin-loppuraportti-perustulokokeilun-tteuttamisvaihtoehdoista>
- Raitano, M., Busilacchi, G., & Gallo, G. (Eds.). (2021). *The politics and economics of basic income experiments*. Edward Elgar Publishing. Retrieved from: <https://www.elgaronline.com/edcollbook-0a/edcoll/9781839104848/9781839104848.xml>
- Tuovinen, A.-K. (2017). Perustuslainmukainen perustulokokeilu: Perustulokokeilun arviointia perustuslain ja kansainvälisten ihmisoikeussopimusten valossa, Kela. Retrieved from: <https://helda.helsinki.fi/items/e60c60af-5879-41cc-ace4-047b0ca2fdab>



Nordic Economic Policy Review 2026

Learning to Experiment: Practical Lessons from Finland's Two-Year Preschool Experiment

Ramin Izadi and Matti Sarvimäki

Abstract

This paper describes how Finland implemented a nationwide randomised field experiment to evaluate a potential reform of its early childhood education and care (ECEC) system. The two-year preschool experiment involved 37,357 children at 956 centres and was enabled by temporary legislation enacted specifically for this purpose. The trial was designed and executed in close collaboration between researchers, civil servants, and teachers, with two successive governments committing to wait for the results before making policy decisions. We document how the experiment was planned, legislated, and implemented, and analyse the institutional arrangements that made it possible. We highlight the value of sustained co-operation between policymakers and researchers, the benefits of a multidisciplinary research team, and how the experiment both relied on and reinforced Finland's data infrastructure. We conclude by discussing lessons for other countries seeking to embed rigorous evaluation into their policy processes.

Acknowledgements: We would like to thank Morten Visby Krægpøth, Ingeborg Foldøy Solli, Roope Uusitalo, Antti Kauhanen, and participants at the Advancing Policy Through Randomized Experiments, Nordic Economic Policy Review Conference, for helpful comments and suggestions. All remaining errors are our own. We gratefully acknowledge financial support from the Research Council of Finland (#358924, #358946).

1 Introduction

In 2021, Finland launched an exceptionally large and ambitious policy experiment. The initiative reflected a long-standing debate about the role of early childhood education and care (ECEC) and preschool as the foundation for lifelong learning. Among policymakers, a broad consensus had emerged that extending preschool from one to two years could be a promising way to improve children's skills and promote equality of opportunity. In practice, this would mean that preschool would begin at age five rather than six, while children would still start primary school at age seven.

Before implementing such a reform, the government decided to generate evidence on its potential effects. As a result, we implemented a large-scale randomised field experiment designed to assess the impacts of extending preschool on children's academic and socioemotional skills, as well as on educational practices, parental experiences, and the functioning of early education centres. Preparations for the experiment began in spring 2020, and it was launched in autumn 2021. The government-commissioned final report, focusing on effects measured at the start of primary education, was published in January 2026. However, the project is planned to continue far into the future. Skill assessments are planned until the end of comprehensive schooling in the early 2030s, and register-based follow-up will allow long-term analysis of children's eventual educational and labour market outcomes.

The trial covered 37,357 children born in 2016 and 2017 at 956 early education centres in 148 municipalities, corresponding to roughly one-third of all Finnish children in those birth cohorts. To enable the experiment, the Finnish Parliament enacted special temporary legislation that specified the randomisation procedure, mandated municipal participation, and authorised the collection of extensive data. The government also committed to taking the results of this experiment into account when deciding whether to reform preschool. This commitment was subsequently upheld by the succeeding government from the opposite end of Finland's political spectrum, making the initiative an exceptional example of policy continuity across governments from opposing political coalitions.

This paper documents how the experiment was planned, legislated, and implemented. Section 2 provides background on how Finnish policymakers gradually moved from a cautious attitude toward randomisation to accepting rigorous experimentation as a legitimate tool for policy development. We describe how early academic trials and often discouraging discussions with government officials eventually gave way to a change. This account is necessarily somewhat subjective and incomplete, but we hope it will inform and inspire similar developments in other countries.

Section 3 describes our collaboration with the Ministry of Education and Culture in setting up the two-year preschool experiment. The process began with a single sentence fragment in the 2019 government programme of Prime Minister Antti Rinne

(later replaced by Sanna Marin), which stated that the government would “pilot two-year preschool.” We outline how these few words were first translated into a research proposal and subsequently into temporary legislation. Again, we offer a necessarily partial and subjective account of this process, which we hope will be useful for both policymakers and researchers seeking to undertake similar initiatives.

Section 4 details our randomisation process. We discuss the challenges of moving from the simple textbook idea of random assignment to implementing it in a complex real-world policy setting. In our case, the treatment had to be assigned at the centre level, even though children are often moved between centres, and some did not attend centre-based early education at the time of randomisation. These features created complications that required both technical and administrative solutions, which we describe in some detail. We also share several practical lessons learned along the way, such as how re-randomisation procedures can be used not only to improve covariate balance but also to ensure that the experiment remains within budget and administratively feasible. Finally, we discuss how randomisation can be communicated to participating municipalities and parents in ways that maintain transparency and trust. In particular, we emphasise the importance of allowing some degree of non-compliance to facilitate smooth implementation.

In Section 5, we discuss the data collected as part of the experiment. Working closely with colleagues from educational sciences, psychology, and learning analytics, we conducted large-scale assessments of children’s academic and socioemotional skills at ages 5, 6, and 7. Academic skills were measured through direct assessments, in which each child completed a structured session with a teacher using a tablet-based test in ECEC and a computer-assisted test with headphones in primary school. Socioemotional skills were assessed by teachers using standardised rating instruments. These assessments were used to construct six preregistered measures: two capturing academic skills (numeracy and literacy) and four capturing socioemotional skills (task performance, social skills, emotional regulation, and peer relationships).

These data were linked to extensive administrative records covering children, their parents, teachers, peers, schools, and early education centres. For each individual, we observe detailed information on education, labour market outcomes, demographics, and family composition. In addition, our colleagues collected rich survey data that provided a much more detailed understanding of how the treatment was implemented in practice than would have been possible using register data alone. Finally, we assembled new administrative information that should ideally be incorporated into national registers – such as data on which teacher was assigned to which group – but is not yet systematically recorded.

Section 6 describes the nature of the treatment. Using large-scale survey and administrative data collected as part of the experiment, we show that the reform altered children’s learning environments in measurable but modest ways. Treatment

centres had more highly qualified teachers, more homogeneous age groups, and slightly more guided learning time, while overall resources, group sizes, and daily schedules remained largely unchanged.

Section 7 summarises the main findings on children's skills. We show that extending preschool increased participation in formal early education at age five by shifting a small group of children from home care into centre-based provision. For these children, two-year preschool generated sizeable, short-term gains in academic and socioemotional skills, but the persistence of these gains remains uncertain. For the large majority of children, however, the relevant alternative was standard daycare. For this group, extending preschool had no effect on skills at school entry, and the scale of the experiment allows us to rule out even modest policy-relevant impacts with confidence. As a result, average effects at the population level are essentially zero.

We conclude with a discussion of broader lessons from the experiment, emphasising how similar large-scale randomised evaluations could be embedded into policy processes in other countries.

2 Randomised field experiments in Finland

This section places the two-year preschool experiment within Finland's gradual shift toward government-sponsored randomised policy trials. Over the past decade, legal and administrative frameworks have been adapted to support experimentation. While gaps remain, rigorous evaluation is now broadly accepted, and even encouraged, across major political parties. We present a brief discussion of how this shift occurred, highlighting details likely to be useful to researchers, civil servants, and policymakers in contexts where randomisation is still considered problematic.

2.1 Early steps

Around the turn of the millennium, economists in Finland, much like their peers elsewhere, began advocating more forcefully for policy evaluation based on credible research designs. Much of the early push focused on bringing quasi-experimental methods to administrative data and programme evaluation, but there were also early calls to run government-sponsored randomised experiments (e.g., Hämäläinen and Uusitalo 2005).

For much of the 2000s and early 2010s, however, progress was limited, and randomised policy trials were often dismissed as constitutionally problematic. This objection appeared to be largely based on a cursory reading of the Constitution's equality clause rather than on careful legal analysis. Instead, ministries and agencies tended to favour regional pilots selected administratively or via applications by willing

participants. Although these pilots also led to differential treatment for citizens depending on where they lived, they were not generally perceived as raising the same constitutional concerns as randomisation. This interpretation of the law appeared illogical and was deeply frustrating for the research community, particularly given that such designs were unlikely to yield credible evidence.

While most policymakers remained cautious, a few government agencies took part in relatively light-touch randomised experiments. In the late 1990s, the Ministry of Labour and public employment offices collaborated with psychologists at the Finnish Institute of Occupational Health on a series of small-scale randomised field experiments evaluating job-search training programmes (Vuori et al. 2002; Hämäläinen et al. 2008). In 2012, the Finnish Tax Administration participated in a study that randomised whether firms received letters providing information about VAT rules (Kosonen and Ropponen 2015). In 2014, Finnish Customs partnered with researchers on an experiment in the used-car import sector, in which randomly selected potential importers were informed that odometer readings would be verified through inspections (Harju et al. 2020). Meanwhile, academics continued to implement RCTs within their own research programmes. A prominent example is the KiVa antibullying programme, evaluated in a large RCT in 2007–2009 and later scaled nationally (Kärnä et al. 2011; Salmivalli and Poskiparta 2012). Another example is an information intervention in high schools in 2011–2012 (Pekkala Kerr et al. 2020).

2.2 The basic income experiment

The launch of the basic income experiment marked an important step toward governmental involvement in randomised trials. The aim was to examine whether simplifying the benefit system and eliminating benefit withdrawal by raising income would increase employment. The experiment was part of Prime Minister Juha Sipilä's government programme and became the first randomised experiment to be implemented under a dedicated piece of legislation. It also led the Constitutional Law Committee to spell out the constitutional preconditions for such experiments.^[6] The Ministry of Justice later codified these principles in a 2020 guide for drafting legislation enabling experiments. The topic of this paper, the two-year preschool experiment, is the second implemented via this legislative route, which is why we discuss the first – the basic income experiment – in some detail.

The final design of the basic income experiment was narrower than the one proposed

6. In its statement (PeVL 51/2016 vp), the Committee deemed the basic income experiment acceptable because it (i) pursued a legitimate objective (producing evidence for future policy decisions); (ii) was authorised by a specific, precise, and time-limited act defining the target population, scope, and deviations from ordinary law; (iii) respected equality and proportionality through randomisation; (iv) did not reduce participants' statutory benefit levels; and (v) ensured regulatory clarity and transparency.

by the researchers commissioned to prepare a proposal for it (Kangas and Pulkka 2016). As a result, it was widely viewed as ill-suited to evaluating basic income as a comprehensive reform (e.g., De Wispelaere et al. 2018; Hämäläinen and Verho 2022). Nevertheless, it created a clean research design for assessing how long-term unemployed individuals respond to stronger financial work incentives and for studying the welfare effects of streamlining the welfare system.

For the experiment, 2,000 individuals on the minimum unemployment benefit were randomly assigned to a tax-free basic income of €560 per month for 2017–2018. The comparison group comprised the remaining eligible beneficiaries (about 175,000 individuals). The basic income equalled the minimum unemployment benefit, meaning that disposable income while unemployed remained largely unchanged. Importantly, however, the benefit did not depend on earnings and job-search requirements. As a consequence, participation tax rates for full-time work fell by an average of roughly 23 percentage points.

The dramatic reduction in the participation tax rate had little effect on employment. During the first year, the estimated impact on annual days employed was statistically insignificant, with a point estimate of only 1.5 days (95% confidence interval: –2.3 to 5.4 days) relative to a baseline of 49 days in the control group (Verho et al. 2022). In the second year, the estimated effect was 6.6 days and statistically significant. However, the research design for this second year was compromised by another policy reform that affected only the control group.

On the other hand, survey evidence indicated greater life satisfaction, lower perceived stress, and fewer bureaucratic hassles for recipients (Tuulio-Henriksson and Simanainen 2020). However, the low overall response rate and the substantial difference in response rates between the treatment (31%) and control (20%) groups warrant caution in interpreting these differences as causal effects. These findings are nevertheless consistent with a follow-up study using administrative health registers, which found that receiving basic income reduced the use of psychotropic medication by 8–11% (Hämäläinen et al. 2025).

2.3 Towards a culture of experimentation

The basic income experiment was part of a broader ambition of the Sipilä government to foster a “culture of experimentation”. The government’s strategic programme emphasised experimentation and established a dedicated Policy Experimentation Unit in the Prime Minister’s Office to coordinate and support pilot projects. Under this agenda, a large number of trials were launched in several policy areas.^[7]

While this approach reflected strong institutional enthusiasm for experimentation,

7. The government launched at least two websites—kokeilunpaikka.fi and kokeilevasuomi.fi—which listed experiments and pilots and provided guidance for conducting them. Both have since been taken down. The last archived version available through the Wayback Machine lists 904 experiments, of which 151 were reported as having received funding (<https://web.archive.org/web/20231204065907/https://www.kokeilunpaikka.fi/fi>).

systematic evaluation was often omitted. According to an official review of the experimentation agenda, many projects labelled as experiments did not meet the criteria of genuine experimentation: they lacked clear hypotheses, comparison groups, and plans for scaling or policy learning (Antikainen 2019). As such, despite the broad rhetoric of experimentation, few initiatives were implemented in ways that could generate robust evidence. Beyond the basic income experiment, we are aware of only two other randomised field experiments conducted during this period in collaboration with government agencies: (i) an experiment with the Finnish Tax Administration that first identified suspected landlords using administrative register data and then randomised them into receiving informational letters about rental income taxation and enforcement (Eerola et al. 2025), and (ii) a trial of a new approach to immigrant integration services with the Ministry of Economic Affairs and Employment (Karinen et al. 2024; Pesola et al. 2025). Both were implemented independently of the Prime Minister's Office and the experimentation agenda, but were in line with the increasing openness to randomised experiments.

The subsequent Marin government made less fuss but continued to develop government-sponsored RCTs. Most notably, it launched the Two-Year Pre-primary Education Trial that is the subject of this paper. Other examples include a collaboration with the Ministry of Economic Affairs and Employment, in which randomly selected entrepreneurs received a subsidy for hiring their first employee (Einiö and Nivala 2025), and a collaboration with the Prime Minister's Office and the Ministry of Justice, in which text messages were sent to randomly selected young voters to encourage electoral participation (Hirvonen et al. 2024).

Government-sponsored randomised trials have continued under Prime Minister Petteri Orpo's government. A recent example is an experiment conducted in collaboration with the Ministry of Social Affairs and Health, in which access to a digital nurse reception and chat service was randomised among roughly 170,000 residents to evaluate how such services affect the use of primary and emergency care (Haaga et al. 2025). Similarly, the City of Helsinki launched a trial in which second-grade classrooms were randomly assigned to receive an additional teacher for four weekly Finnish lessons (Holvio et al. 2024).

3 Setting up the two-year preschool experiment

The origins of the two-year preschool experiment trace back to the negotiations to form a government after the general election in spring 2019. In the run-up to the elections, extending preschool featured in the platforms of most major political parties.^[8] The elections resulted in a left-leaning coalition that adopted an ambitious agenda for educational reform. Its main initiative was to extend compulsory schooling from 16 to 18, while the plan to expand preschool was reduced to a brief line in the government programme stating that the government would “pilot two-year preschool.” Although this reference might have seemed minor at the time, it provided the political mandate that ultimately enabled a nationwide randomised experiment.

3.1 From mandate to proposal

Once the new government was formed, preparations for implementing its programme began. As part of this process, the Ministry of Education and Culture invited representatives from the Aalto Economic Institute in February 2020 to discuss how the pilot could be designed and evaluated in practice. These initial meetings developed into a series of detailed discussions between researchers and civil servants. The ministry subsequently requested a formal research proposal, which we submitted as (Izadi et al. 2020).

The proposal set out a preliminary research design, a power calculation, and an initial plan for measuring children’s skills. It also explained why strict compliance was neither necessary nor desirable. We aimed to avoid situations where participation would clearly conflict with a child’s best interests and to allow reasonable flexibility when logistical considerations or parents’ beliefs about the suitability of two-year preschool strongly conflicted with the results of the randomisation. We also stressed that non-compliance reduces the statistical power and representativeness, underscoring the importance of making compliance as easy and pleasant as possible.

We further noted that the standard way of conducting pilots at the ministry, based on voluntary participation, detracts from the credibility of causal inference. By contrast, random assignment would ensure comparability between treatment and control groups and provide a sound basis for evaluating the policy’s effects. To maximise statistical power, we proposed assignment at the centre level rather than at the municipality level.

8. The National Coalition Party, the Greens, the Left Alliance, and the Swedish People’s Party all proposed moving the starting age for preschool from six to five. The Christian Democrats advocated voluntary but free preschool for five-year-olds, while the Social Democratic Party pledged to decide on expanding preschool based on the results of future trials. The remaining two major parties adopted different positions: the Centre Party proposed establishing a new type of school for five- to seven-year-olds, and the Finns Party made no reference to ECEC in its programme. The 2019 government was a coalition between the Social Democratic Party, the Green League, the Left Alliance, the Swedish People’s Party, and the Centre Party, led by Prime Minister Antti Rinne. Following Rinne’s resignation later that year, the coalition continued under Prime Minister Sanna Marin with the same party composition and programme.

3.2 From proposal to legislation

Following the research proposal, the Ministry of Education and Culture, with assistance from the research team, prepared a government bill for a fixed-term act to enable the experiment. The draft was circulated for public comments in summer 2020 via the standard consultation portal and received 56 written statements.^[9] The government then submitted its proposal (HE 149/2020) to Parliament, introducing a temporary act and minor amendments to the Early Childhood Education Act. The bill specified the randomisation protocol, defined the 2016–2017 birth cohorts as the target group, and authorised the collection and linkage of evaluation and administrative data for research purposes.

The proposal proceeded through the ordinary committee process. The Constitutional Law Committee (PeVL 37/2020) concluded that the experiment was compatible with the constitutional provisions on equality and the right to education because participation was voluntary, the intervention was time-limited, all children retained access to standard one-year preschool, and the differential treatment rested on an objective and acceptable reason, i.e., producing evidence to inform future policy. The Committee also emphasised the need for clear statutory definitions of participants' rights and obligations and for adequate data-protection provisions.

The Parliament's Education and Culture Committee (SiVM 11/2020) similarly endorsed the bill with minor technical amendments. Its report highlighted the value of evaluating major education reforms before permanent adoption, supported randomisation, and stressed the importance of explicit provisions enabling rigorous evaluation alongside safeguards for personal data. Finally, in the plenary debate (PTK 154/2020), members expressed broad support across party lines, with the only critical remarks coming from the Finns Party. Parliament adopted the law, and the Temporary Act 1046/2020 entered into force on 23 December 2020.

4 Experimental design

This section describes the design and implementation of the experiment and explains why randomised controlled trials offer the most reliable basis for assessing policy impacts. We also discuss several challenges we encountered and how we addressed them.

9. The call for consultation and all statements are available at www.lausuntopalvelu.fi/Fi/Proposal/Participation?proposalId=d8762499-96b3-452b-95a0-05ade359b178.

4.1 Rationale for randomisation

The primary aim in designing the experiment was to assess how implementing legislation that would extend preschool to two years in Finland would affect children's development. In practice, this required comparing children who attended two years of preschool with similar children who remained in the current system of regular daycare (or home care) and then entered the existing one-year preschool programme. Without an experimental design, such comparisons would very likely be biased, as families who voluntarily choose longer preschool are likely to differ systematically from those who do not.

Random assignment solves this problem. By randomly allocating children to either the new two-year preschool or the standard one-year system, the trial ensured that, on average, the groups were equivalent in all relevant respects before the intervention. Any subsequent differences in outcomes can, therefore, be attributed to the policy itself rather than to pre-existing differences between families.

4.2 Randomisation design

The goal was to build a credible and administratively feasible research design while minimising the burden on municipalities and families. Treatment assignment needed to be implemented in a way that ensured all children were treated fairly and lawfully. The final design was a randomised field experiment in which municipalities or centres were selected by lottery.

4.2.1 Practical constraints that shaped the design

1. The research design had to satisfy several practical and institutional requirements:
2. The intervention could only be implemented at the centre level, but each child's treatment status needed to be defined. Parents had to be personally informed whether their child was assigned to the treatment or control group and asked to enrol their child accordingly.
3. The target population encompassed all children born in 2016 and 2017, including those not enrolled in centre-based care.
4. Only centres offering both preschool and early childhood education were deemed eligible so that children could remain in the same centre throughout the day. Arrangements in which preschool was provided in schools were, therefore, excluded from the experiment.

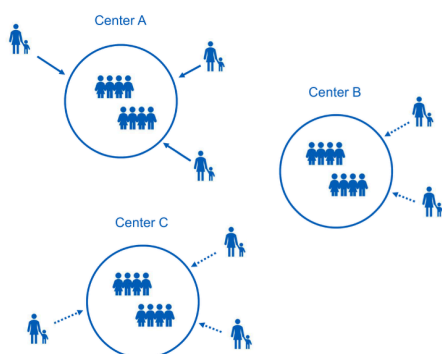
The design needed to accommodate municipalities of all sizes, including those with only a few centres.

4.2.2 Solutions to 1–3

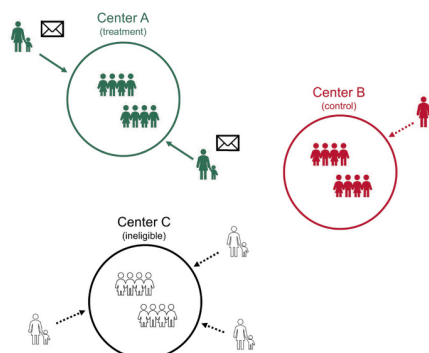
Every child had to be linked to a specific early childhood education centre before randomisation so that treatment status could be determined unambiguously. For enrolled children, the link was determined by their current centre. The remaining 15% were not yet enrolled but were of particular interest to policymakers, who sought to increase participation. To balance participation with capacity constraints, we first assigned each unenrolled child to their nearest centre based on straight-line distance. Municipalities then selected as many children as each centre could accommodate. Children not selected were excluded from the experimental sample, as were those assigned to ineligible centres. Randomisation was conducted at the centre level, and the treatment group comprised children assigned to treatment centres. Similarly, the control group consisted of children assigned to a control centre. Figure 1 illustrates this approach.

Figure 1. Experimental design in large municipalities

Panel A: Pre-randomisation



Panel B: Post-randomisation



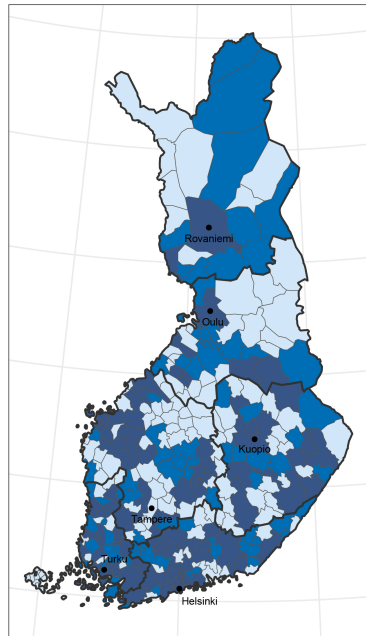
Note: Illustration of the experimental design in municipalities with more than five eligible daycare centres. Children are assigned to the centre they are enrolled in or the centre closest to their home. Eligible daycare centres are randomised into treatment and control groups, and children are invited to participate.

4.2.3 Solution to 4

Finnish municipalities vary greatly in size, and some have only a handful of eligible centres. Because participation in the treatment involved a fixed cost at the municipal level, it would have been unreasonable to ask municipalities to bear this cost if they had only a few treatment centres. At the same time, nationwide coverage was deemed important. To address this, we divided municipalities into two groups. In large municipalities with five or more eligible centres, randomisation occurred between centres within each municipality, as discussed above. Roughly 40% of centres were assigned to treatment and 60% to control. In small municipalities with one to four eligible centres, the entire municipality was randomised to treatment or control. Treatment municipalities implemented two-year preschool in all eligible centres, while control municipalities continued with the existing system. Panel (a) of Figure 2 shows the participating municipalities on the map. It also shows that many municipalities lacked eligible centres, which excluded them from the experiment. (These are municipalities where preschool takes place in primary schools.) Additionally, to stay within budget, some large municipalities had to be excluded from the trial. Panel (b) shows which randomly sampled large municipalities were included in the trial and which were excluded. To ensure broad regional coverage, sampling was stratified geographically by Regional State Administrative Agency (AVI) area and municipality size.

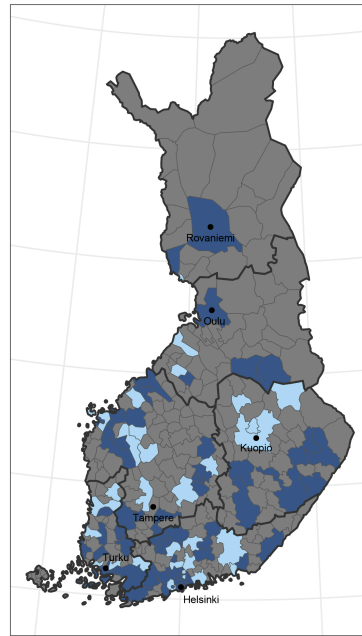
Figure 2. Sampling frame

Panel A: Municipalities with eligible daycare centers



Large munis (4+ centers) Small munis (1-4 centers) No eligible centers

Panel B: Large municipalities' sampling



In sample Out of sample

Lessons Learned: Making Better Use of Oversubscription

Many centres were capacity-constrained when admitting unenrolled children. In retrospect, rather than asking municipalities to select children from a pre-assigned nearest-centre list, we could have (i) collected only seat capacities from municipalities, and then (ii) randomised unenrolled children *individually* to fill those seats. Children who could not be accommodated would have formed a natural control group. This alternative would have increased effective sample size and statistical power without adding administrative burden.

4.3 Randomisation

Randomisation was based on administrative data from spring 2021 covering children born in 2016 and their families. To avoid unnecessary organisational disruption, centres randomised to treatment in 2021 (for the 2016 birth cohort) retained their status the following year when the 2017 cohort began the first year of two-year preschool. A foreseeable consequence was that some families of children born in 2017 could be sorted into treated centres. We anticipated this risk at the design stage and committed to addressing any resulting selection empirically in the analysis (see balance results below).

4.3.1 Data

Constructing the data infrastructure required to implement the experiment was itself a major policy innovation. Finland's renowned centralised administrative registers did not yet cover early childhood education, so creating a national sampling frame required building one from scratch. Working with the Ministry of Education and Culture, we compiled enrolment lists from every municipality and verified them locally. This collaboration effectively transformed a fragmented set of local records into Finland's first national ECEC register. The temporary legislation authorising the experiment also allowed us to use children's social security numbers directly for randomisation, which greatly facilitated the process of informing parents about their child's treatment status. These personal identifiers also made it possible to link the enrolment data to the national population register, allowing us to assign nearby centres to unenrolled children and ensure complete coverage. The entire sampling process was carefully documented and archived in the National Archives for future replication and research use.

4.3.2 Re-randomisation

We implemented a stratified re-randomisation procedure as per Morgan and Rubin (2012) and Johansson et al. (2021). Specifically, we simulated one million potential treatment allocations, retained the 2,000 that achieved the best balance across key covariates, verified that these allocations were not deterministic, and randomly selected the final assignment from among them. The procedure served three purposes. First, it ensured good covariate balance between treatment and control centres with respect to characteristics observed at the time of randomisation. Second, it allowed us to calculate exact t -values, as we know all possible treatment assignments that could have been selected. Third, and most importantly, it enabled us to impose a strict budget constraint: because government reimbursements to municipalities depended on centre size and the expected enrolment of unenrolled children, we excluded allocations that would have exceeded the fixed €30 million budget.

4.4 Assessing randomisation balance

Random assignment should, in expectation, create treatment and control groups that are identical in average terms. In practice, chance variation can lead to small imbalances, which is why randomised evaluations typically begin by testing whether background characteristics are balanced between groups. In our case, the fact that centres were randomised only during the first year of the experiment provides an additional rationale for checking balance.

Table 1. Covariate balance by cohort

	2016 Cohort		2017 Cohort	
	Control	Trt – Ctrl	Control	Trt – Ctrl
Mother's earnings (€)	23,253	167	23,617	992**
	(294)	(399)	(282)	(446)
Father's earnings (€)	43,372	-14	44,064	928
	(621)	(863)	(621)	(876)
Mother's education (years)	13.33	0.10	13.34	0.15**
	(0.05)	(0.07)	(0.05)	(0.07)
Father's education (years)	12.83	0.10	12.86	0.10
	(0.05)	(0.07)	(0.05)	(0.07)
Single parent	0.20	0.00	0.20	0.00
	(0.01)	(0.01)	(0.01)	(0.01)
Number of siblings	1.59	-0.03	1.65	-0.09**
	(0.04)	(0.04)	(0.04)	(0.04)
Girl	0.49	0.00	0.48	0.01*
	(0.00)	(0.01)	(0.00)	(0.01)
Mother's domestic language	0.84	0.02*	0.83	0.02**
	(0.01)	(0.01)	(0.01)	(0.01)
Father's domestic language	0.84	0.01	0.83	0.02**
	(0.01)	(0.01)	(0.01)	(0.01)
Child's domestic language	0.88	0.01	0.87	0.02**
	(0.01)	(0.01)	(0.01)	(0.01)
Immigrant parents	0.13	-0.01	0.14	-0.02**
	(0.01)	(0.01)	(0.01)	(0.01)
Foreign-born child	0.03	0.00	0.03	0.00
	(0.00)	(0.00)	(0.00)	(0.00)
Immigration age (years)	1.98	-0.08	2.05	0.09
	(0.07)	(0.11)	(0.06)	(0.11)
Exact Wald -value				
Separate regressions	-	0.409	-	0.040
Joint regression	-	0.562	-	0.084
Asymptotic Wald -value	-	0.629	-	0.073
Sample size	10,899	18,932	10,461	18,305

Note: Standard errors in parentheses. *, **, ***.

Table 1 reports the balance checks. For the 2016 cohort, randomisation produced highly comparable treatment and control groups. For the 2017 cohort, some modest but statistically significant differences emerged: children in the treatment group tended to come from slightly higher-income and more educated families and were more often native speakers of Finnish, Swedish, or Sámi. These minor deviations likely arose because treatment and control centres retained their status between cohorts. Families of children born in 2016 could not influence their assignment, as randomisation took place after their preschool enrolment decisions. In contrast, families of 2017-born children could, in principle, move their children to a treated or control centre before their children's treatment status was determined.

Although such selection could threaten causal interpretation, its magnitude appears limited. Our research design also allows us to assess this possibility directly. Most importantly, we can compare estimates using data on the 2016 and 2017 birth cohorts separately. In addition, we can examine how sensitive the results are to the inclusion of observable characteristics in the specifications. In practice, these analyses suggest that potential selection bias in the 2017 cohort is unlikely to affect our conclusions.

5 Children's skills assessments

5.1 Purpose and rationale

Once the research design was finalised and randomisation completed, the next challenge was to measure child development to examine how starting preschool a year earlier affected the foundations for learning and social participation at school entry. To answer this question credibly, the study needed precise, developmentally appropriate measures that could be administered at scale across almost a thousand early education centres.

While economists could design the experiment and analyse the results, assessing five-year-olds required a different kind of expertise, and the research group was extended to include early childhood education specialists, developmental psychologists and learning analysts. This multidisciplinary collaboration was essential to ensure that the chosen instruments captured genuine developmental progress.

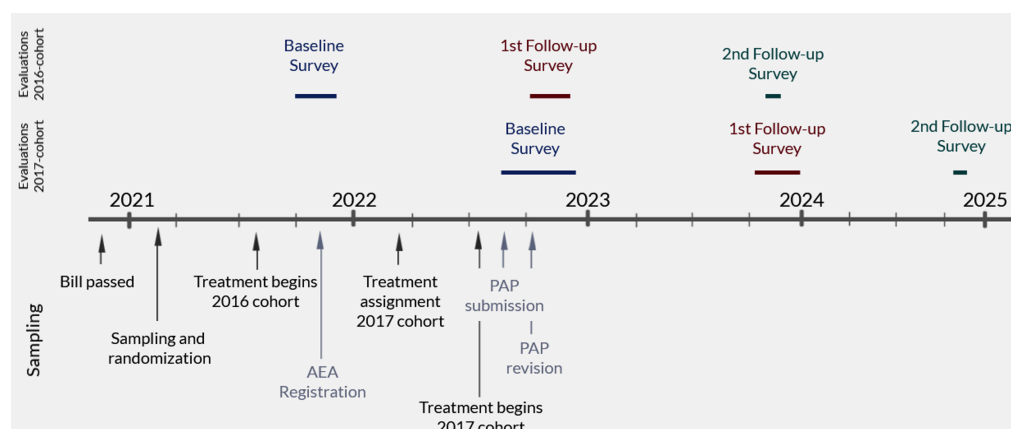
The assessment design sought to cover the full range of abilities within each age group. To follow children over time, each test wave became slightly more demanding, with overlapping items linking the results into a consistent developmental scale. In short, the assessments were designed to provide reliable evidence on how children learn and grow during the preschool years, and to do so in a way that could inform national policy decisions about early education.

5.2 Assessment design and timeline

Children's skills were assessed three times between the ages of five and seven to capture both immediate and lasting effects of the two-year preschool programme. The assessments took place in the autumns of 2021–2024 and followed the progression of the two birth cohorts included in the experiment. Each wave used the same core domains, enabling consistent tracking of changes over time.

The first assessment served as a baseline. It was conducted when treatment-group children began preschool at age five and control-group children remained in daycare or at home. The second round took place a year later, when all treatment children had completed their first preschool year, and most control children had just entered the regular one-year preschool. The final assessment was carried out at school entry, when all children were seven years old and enrolled in first grade. Figure 3 shows the timeline.

Figure 3. RCT Timeline



5.3 Assessment at scale

Assessing more than 30,000 children across the country required a system that was both uniform and practical for everyday use. The solution was to rely on teachers to conduct the assessments and to use the *VILLE* learning analytics platform developed by the Turku Research Institute for Learning Analytics at the University of Turku. The platform enabled standardised administration, automatic scoring, and secure data transfer, substantially reducing both workload and error compared with paper-based testing.

Figure 4. Data collection

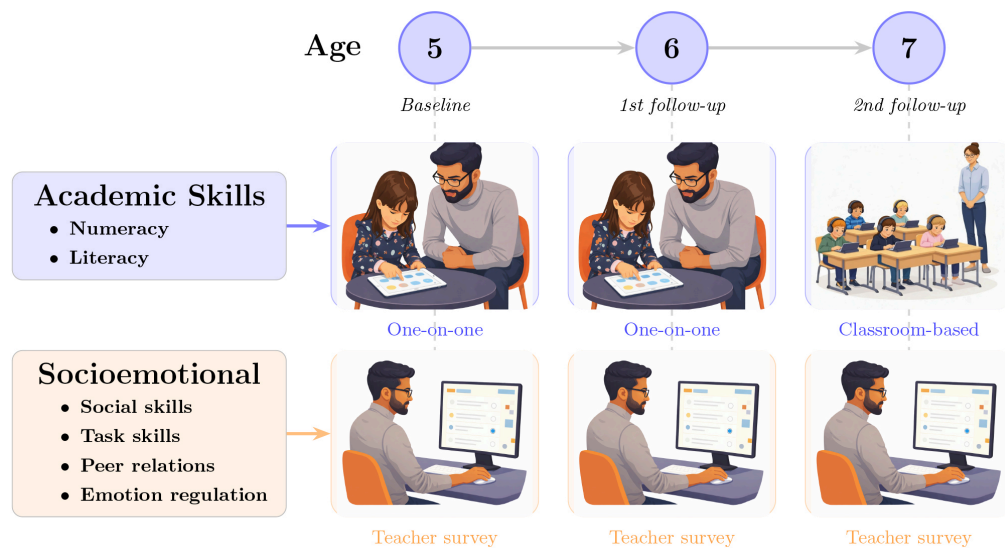


Figure 4 summarises the data collection method. Teachers supervised the sessions in familiar classroom settings. In preschool, assessments were conducted one-on-one with a tablet, while in first grade, entire classes completed them during two sessions under teacher supervision using tablets/laptops and audio instructions from headphones. Each child received a personal login linked to their national identifier, enabling seamless linkage to register data for later analysis. To make participation easy, even for young children, the system featured a picture-based sign-in interface that allowed them to log in independently.

Before testing, teachers completed short online training modules and received detailed written and video instructions. The platform provided immediate quality checks, time stamps, and error logs, ensuring consistent implementation across municipalities. Teachers recorded socioemotional ratings directly within the same system, making *VILLE* a single digital hub for all parts of the assessment.

This digital infrastructure turned what could have been a logistical bottleneck into a strength. It enabled uniform, high-quality measurement across hundreds of centres while creating a permanent, linkable dataset that will support long-term follow-up studies.

5.4 What was measured

The assessments covered both socioemotional and academic skills using instruments developed or adapted in collaboration with Finnish and international experts. Table 2 lists all instruments and their sources.

Table 2. Assessment instruments

Acronym	Source	Description / Notes
Social-Emotional Skills (Teacher Rated)		
BSRS	Onatsu-Arvilommi and Nurmi (2000)	Measures task avoidance and engagement strategies.
MASCS	Merrell (1993)	Short version assessing prosociality; excluded in grade 1.
CBRS	Bronson et al. (1990)	Short version assessing self-regulation.
SDQ	Goodman (1997)	Five subscales (emotional symptoms, conduct, hyperactivity, peer problems, prosocial behaviour).
SSES	OECD (2021)	Measures curiosity and happiness (three items each).
Language Skills		
PPVT	Dunn and Dunn (2007)	Picture vocabulary assessment.
ARMI	Lerkkanen et al. (2006)	Finnish early learning battery covering phonological awareness, letter knowledge, and reading skills.
Numeric Skills		
–	Project-specific	Counting numbers (two tasks at age 5, four at age 6); quantity comparison (45-second limit per task); quantity production (quantities 3, 7, 13, 21); arithmetic operations (four tasks at age 5, six at age 6).
General Cognitive Skills		
WJ	Woodcock and Johnson (1977)	Spatial relations test with 3-minute time limit.

Note: All measures were administered through an online platform. Teacher ratings use Likert-type scales. Project-specific measures were developed and validated for this study. Time limits are indicated where applicable. Acronyms: BSRS = Behavioural Strategy Rating Scale; MASCS = Multisource Assessment of Social Competence; CBRS = Child Behaviour Rating Scale; SDQ = Strengths and Difficulties Questionnaire; SSES = Survey on Social and Emotional Skills; PPVT = Peabody Picture Vocabulary Test; ARMI = Finnish early learning assessment battery; WJ = Woodcock-Johnson Spatial Relations Test.

5.4.1 Academic development

The academic assessments were designed to capture the skills that form the foundation of early learning: understanding quantities, manipulating numbers, recognising letters and words, and linking sounds to text. They drew on established Finnish early-learning batteries (ARMI) and international measures. Each task was piloted and calibrated to span the entire skill distribution. The most basic items asked children to identify letters or count small sets of objects, while more advanced ones required reading simple words or solving short arithmetic problems. As children grew older, the tasks became more challenging, with overlapping items anchoring results across waves to ensure comparability. Items were presented as short, game-like tasks that children could complete within 20–25 minutes. Automated scoring provided precise and comparable indicators for every child.

5.4.2 Socioemotional development

To capture children's behavioural and emotional adjustment, teachers completed standardised rating scales covering self-regulation, task focus, empathy, co-operation, and emotional well-being. The instruments combined well-validated international measures from several sources (see Table 2). Teachers entered their responses directly into the *ViLLE* platform using each child's individual account, ensuring consistency and secure linkage to academic outcomes.

5.5 Pre-analysis plan: Defining outcomes and analysis in advance

We filed a pre-analysis plan outlining the main research questions, outcome measures, and statistical methods in advance. The plan identified six primary outcomes: two for academic skills (numeracy and literacy) and four for socioemotional development (task skills, social skills, peer relations, and emotion regulation). Each outcome combined several underlying test items or teacher-rated scales into a single, interpretable measure. These definitions were agreed upon and registered before the first child was assessed at age six, ensuring that the outcome categories were fixed in advance rather than shaped by the data.

The value of this approach is to improve credibility. A pre-analysis plan does not alter the results, but it strengthens confidence that the findings reflect what the experiment was designed to test rather than what later appeared interesting. Our plan also specified the main estimation strategy, subgroup analyses, and robustness checks to be conducted. Pre-analysis plans are increasingly standard in experiments because they reduce the scope for selective reporting and make analytical decisions transparent. They serve both scientific and policy purposes by facilitating replication, guiding interpretation, and signalling that the evaluation followed a clearly articulated design rather than post hoc choices.

5.6 Participation and data quality

Assessments covered all children born in 2016 and 2017 who were enrolled in participating treatment and control centres, with families retaining the right to opt out. Participation was high with response rates ranging from 83% to 91%, depending on the cohort and skill domain. Table 3 reports detailed figures by age and cohort. The median completion time per child was 22 minutes for the cognitive tasks and 8 minutes for the teacher evaluations. Children not enrolled in any centre at the time of data collection, primarily five-year-olds in home care, were not assessed. High completion rates were maintained through careful field coordination. A centralised call centre staffed by trained research assistants monitored real-time completion data from the *VILLE* platform and followed up systematically with reminders to centres with unfinished assessments. This approach, combined with digital administration and standardised procedures, minimised missing data and ensured consistent quality across municipalities.

Attrition was modest and largely unrelated to treatment assignment. The main reasons for missing data were relocation outside participating municipalities or temporary absences during the testing window. From age six onward, when both treatment and control groups attended mandatory preschool or school, differential attrition effectively disappeared.

Table 3. Participation in Child Skill Assessments by Age, Cohort, and Domain

Age group	Skill domain	2016 cohort		2017 cohort	
		N	Coverage (%)	N	Coverage (%)
Age 5	Literacy	16,401	91.3	15,197	84.7
	Numeracy	16,373	91.2	15,172	84.5
	Socioemotional	16,200	90.2	14,931	83.2
Age 6	Literacy	19,854	87.6	19,283	87.5
	Numeracy	19,816	87.4	19,274	84.5
	Socioemotional	19,226	84.8	18,638	84.6
Grade 1 (Age 7)	Literacy	31,492	91.9	30,512	94.3
	Numeracy	30,324	88.5	29,361	90.7
	Socioemotional	29,983	87.5	29,120	90.0

5.7 Building a long-term research asset

Although the current analyses focus on short-term effects, the data were designed to serve as the foundation for a long-term research asset. Recognising the scale of the effort, both researchers and policy partners saw the project as an opportunity to build a national resource: a longitudinal dataset on early childhood development that combines rich skills assessments with administrative records. Importantly, the subsequent government incorporated this objective into its policy agenda and committed to continuing the skill assessments through the end of comprehensive education. As a result, the value of the data will extend far beyond the immediate experiment, enabling future research on how early skills shape later outcomes.

Lessons Learned: Putting the Field into Field Experiments

Running an experiment in real classrooms means stepping away from the desk. Members of our team visited daycare centres, watched the new structured morning sessions, and even helped run assessments during pilot rounds. These visits offered a glimpse of how the reform took shape beyond the spreadsheets and gave context to what the numbers would later show. We also held seminars and workshops with teachers, principals, and municipal coordinators across the Helsinki region to hear how the two-year preschool worked in practice.

6 Treatment

To interpret the estimated effects on children's skills, it is essential to understand how the experiment altered the everyday experiences of children participating in the two-year preschool relative to the existing system. To document these changes, the research team conducted large-scale surveys and collected and analysed administrative records, as well as national and local curricula. As a result, we can describe the treatment in considerably greater detail than is typically possible in policy evaluations. We next discuss these findings.

6.1 Legislation

The intervention advanced the start of preschool by one year, so that children in the treatment group attended preschool at ages five and six instead of only at six. The *Act on the Two-Year Preschool Trial* (1046/2020) defined the target population,

randomisation, and implementation rules. Participation was mandatory in principle: families were expected to ensure that their child achieved preschool learning goals, either by attending the assigned programme or by other means.

The new two-year preschool followed a national curriculum prepared by the Finnish National Agency for Education and implemented by municipalities. Preschool was free of charge for four hours each morning during the school year, with optional afternoon daycare subject to income-based fees. Children living more than five kilometres away were entitled to free transportation.

6.2 Curricular content and pedagogical approach

The two-year preschool curriculum built directly on Finland's existing framework for six-year-olds. It remained play-based and child-centred, and in most respects closely resembled the early childhood education and care (ECEC) curriculum followed in the control group. The main goals were similar to those in ordinary daycare, but the two-year preschool curriculum placed somewhat greater emphasis on purposeful group activities led by preschool-qualified teachers. Daily sessions typically combined guided play with early literacy and numeracy tasks. Teachers were encouraged to integrate learning into play, and classroom-style teaching was virtually absent.

Responsibility for implementation followed Finland's three-tier governance model. The National Agency for Education set national objectives; municipal education boards translated them into local curricula; and centres developed their own pedagogical plans. This structure promoted local ownership but also introduced variation across municipalities. Some centres emphasised academic readiness, others focused more on social interaction and emotional growth, reflecting differences in local priorities and staff expertise.

6.3 Changes in educational inputs

The experiment altered children's learning environments in measurable ways. Survey and administrative data reveal systematic differences in teacher qualifications, group composition, and daily time use between treatment and control centres. Table 4 summarises the main contrasts.

Table 4. Changes in educational inputs: Treatment (preschool) vs. Control (daycare)

	Treatment	Control	Difference
	(Preschool)	(Daycare)	(T – C)
A. Highest qualification of teacher(s) in group			
Bachelor's or Master's degree in ECEC (%)	65.4	44.3	+21.1
Bachelor's in social pedagogy/social services (%)	25.1	40.5	15.4
No tertiary degree in ECEC (%)	9.5	15.2	5.7
B. Staffing			
Number of teachers in group (avg.)	1.32	1.23	+0.09
C. Child group's age composition			
Only 5-year-olds (%)	45.3	19.3	+26.0
5–6-year-olds (%)	21.6	13.9	+7.7
5-year-olds and younger children (%)	27.3	53.0	25.7
5–6-year-olds and younger (%)	5.8	13.7	7.9
D. Weekly time use			
Guided activities (hours/week)	7.6	6.5	+1.1
Routines (hours/week)	12.0	14.0	2.0

Note: Differences are expressed in percentage points (pp) or hours. Routines include dressing, meals, rest, and similar activities.

Teacher qualifications improved in treatment centres as a direct consequence of the reform. Owing to randomisation, there were no systematic differences between treatment and control centres prior to the experiment. During the experiment, however, treatment centres had about a six-percentage-point lower share of children taught by staff without any formal qualifications. The composition of qualified teachers also differed between treatment and control groups: children in the treatment group were 21 percentage points more likely to have a teacher with a university degree in early childhood education and 15 percentage points less likely to have a teacher with a degree in social pedagogy from a university of applied sciences. These differences reflected both new hires and the reshuffling of teachers within municipalities. Several municipalities reported reallocating their most experienced staff to treatment classrooms. This response also revealed the limits of national capacity, as only about half of municipalities were able to fill all posts with fully qualified teachers.

Group composition also changed. Treatment centres largely shifted from mixed-age daycare groups (ages three to five) to age-specific preschool groups, typically composed of five-year-olds only. Children were about 26 percentage points more likely to have only same-age peers, enabling a more uniform pace for activities and reducing the time spent assisting younger children. Instructional time increased modestly: teacher time-use surveys indicate roughly one additional guided session per week, or about a 15–20% rise, alongside a two-hour weekly reduction in routine activities such as dressing, eating, and rest. The overall length of the day remained unchanged.

Aside from these dimensions, most features of provision remained stable. Group sizes and child–teacher ratios did not change, daily schedules continued to alternate between play, meals, and rest, and centres relied on the same facilities and materials as before. In sum, the reform primarily reallocated existing resources toward more qualified teachers, more homogeneous age groups, and slightly more guided learning time. These adjustments should be understood as an integral part of the treatment and highlight that a nationwide transition to two-year preschool—if undertaken without a substantial expansion in the supply of qualified teachers—would likely also affect staffing and pedagogical conditions in groups serving younger children.

7 Effects on children's skills

We now turn to the effects on children's skills. This section provides a brief overview of the main findings, drawing on the final evaluation report published (in Finnish) by the Ministry of Education and Culture (Sarvimäki et al. 2026) and a companion academic paper.^[10] We refer readers interested in methodological details and a full presentation of the estimates to these companion papers.

7.1 Intention-to-treat estimates

The primary objective of the experiment was to assess whether a reform replacing voluntary early childhood education and care (ECEC) at age five with a first year of mandatory two-year preschool would better prepare children for primary education. The experiment was designed to approximate as closely as possible the overall effects of such a reform. The scale of the reform was central to this aim: in addition to providing statistical power, it allowed us to study outcomes under realistic conditions in which the estimates capture the combined effects of increased participation, changes in the content of ECEC, and the challenges associated with implementing the reform within the existing system. We begin with a treatment–control comparison that estimates intention-to-treat effects on children's skills. The next subsections trace these differences to changes in participation, content, and implementation.

Table 5 reports estimates of the effects of assignment to the treatment group on indices of academic and socioemotional skills. At age six, children in the treatment group scored slightly higher in academic skills and slightly lower in socioemotional skills than those in the control group. Both estimates are individually marginally statistically significant, with adjusted p-values of 0.069 that account for multiple hypothesis testing across both outcomes. In addition, we reject the joint null hypothesis of no effect on either skill, with a p-value of 0.004.

Taken at face value, the estimates suggest that in the short term, the reform may have involved a trade-off between academic and socioemotional skills. However, we caution against drawing strong conclusions regarding socioemotional outcomes at age six due to the challenges involved in quantifying them. Unlike academic skills, which were assessed using computer-based tests administered directly to children, socioemotional skills were rated by teachers. At age six, treated children were typically evaluated by teachers who had known them for substantially longer than control children, and teachers generally evaluated children from only one experimental condition, either treatment or control. In addition, teachers in treatment centres were directly involved

10. The academic paper was in preparation at the time of writing, but a working version will be available on our websites when this issue of NEPR is published.

in implementing the experiment, which may have shaped their perceptions or reporting in ways that are difficult to disentangle from true skill differences. As a result, observed differences may partly reflect differential familiarity or reference effects rather than underlying changes in socioemotional skills.^[11]

Table 5. Intention-to-treat estimates

	Outcomes			
	Age 6		Age 7	
	Academic	Socioemotional	Academic	Socioemotional
Treatment group	0.031	-0.039	-0.012	0.003
	(0.014)	(0.020)	(0.014)	(0.016)
Cluster p-value	0.032	0.051	0.382	0.846
Exact p-value	0.037	0.059	0.403	0.847
Romano-Wolf p-value	0.069	0.069	0.633	0.847
Wald p-value	0.004		0.666	
N	30,913	29,969	30,940	29,454

Note: This table reports estimates from regressions of children's assessed skills on an indicator for assignment to the treatment group, parental socioeconomic and demographic characteristics, and block-by-cohort fixed effects. Standard errors (in parentheses) are clustered at the level of randomisation (municipality or administrative region). Exact -values are based on permutation tests. Romano-Wolf adjusted -values account for testing two outcomes. The Wald -value tests the joint null hypothesis of no effect on either skill.

A more important observation is that the age-six estimates are statistically significant only because of the high statistical power of the experiment. The point estimate is 0.03 standard deviations for academic skills and standard deviations for socioemotional skills. As a benchmark, effects below 0.05 standard deviations are typically classified as small in the education literature (Kraft 2020). Another way to interpret the magnitude is that an effect of 0.03 standard deviations in academic skills corresponds to roughly two weeks of typical developmental progress at this age.

Most importantly, by the start of primary education, the skills of children assigned to the treatment and control groups had evolved identically. At age seven, the point estimates for average effects are -0.012 and 0.003 standard deviations for academic and socioemotional skills, respectively, and neither estimate is statistically significant. This is not simply an absence of evidence: the estimates are sufficiently precise to conclude that any actual effect is minimal. For the academic skills index, the 95% confidence interval ranges from -0.04 to 0.02 standard deviations. For the

11. These concerns are less pronounced at age seven, when all children were in first grade, typically in mixed classrooms, and rated by teachers who were not directly involved in the experimental intervention and had known the treatment and control children for comparable periods.

socioemotional skills index, we can rule out effect sizes larger than 0.03 standard deviations in either direction.

In our companion work, we also find no evidence that the experiment achieved its stated objective of levelling the playing field at the start of primary education. The intention-to-treat effects remain statistically insignificant, with small point estimates, across subgroups defined by family composition and parental income, education, country of birth, and employment status.^[12] We also examined directly whether two-year preschool compressed the distribution of skills. While there is modest evidence of compression at age six, this pattern does not persist to school entry. If anything, the variance of academic skills at the start of primary education is slightly larger in the treatment group than in the control group. In short, we show that extending preschool did not reduce disparities in skills at school entry.

Interestingly, our findings differ markedly from those of Rege et al. (2024), who study a broadly similar expansion of early education for five-year-olds in Norway. In that study, the estimated effect on a composite index combining all measured skills was 0.12 standard deviations at age six and 0.13 standard deviations at age seven. These estimates are statistically significantly and substantially larger than those observed in our setting. Several factors may account for these differences. First, the relevant counterfactual (standard provision for five-year-olds) may differ between Norway and Finland. Second, the interventions themselves were not identical: the Norwegian reform involved new curricula that appear to have been more detailed and prescriptive than those used in Finland's two-year preschool trial. Third, the scale and mode of implementation differed sharply. The Norwegian experiment was conducted as a pilot involving 30 municipalities, of which 15 chose to participate, with randomisation across 71 units. By contrast, Finland's experiment was implemented in 148 randomly selected municipalities and 956 centres. As a result, our estimates reflect outcomes under conditions in which challenges related to system-wide scaling are fully present, including implementation in municipalities and centres that would not have volunteered to participate. This distinction is important given recent discussions on how treatment effects identified in small-scale pilots may attenuate when policies are implemented at scale (e.g., List 2022).

12. Importantly, we also show that effects are not larger for the 2017 cohort than for the 2016 cohort, alleviating concerns related to rushed implementation or selective sorting into treatment centres.

Lessons Learned: When Null Results Are Informative

A common misconception in interpreting statistical results is that the absence of statistically significant effects implies that a policy has no impact. In many settings, this inference is unwarranted: small or underpowered studies with insignificant results often yield wide confidence intervals and therefore cannot rule out meaningful effects. In our case, however, the situation is different. The large scale of the experiment yields narrow confidence intervals, providing sufficient statistical power to rule out even modest effects. When designing future trials, investing in sample size pays off not only when effects are detected, but especially when tight confidence intervals allow policymakers to conclude that meaningful effects are unlikely.

7.2 Effect on childcare choices

The intention-to-treat estimates discussed above reflect both how the experiment altered participation across different forms of childcare and the relative effectiveness of two-year preschool compared with the existing alternatives. To better understand these findings, we first document how the experiment affected childcare arrangements and then discuss the implications for interpreting the effects on children's skills.

Figure 5. Childcare choices through ages 4-6. Treatment N = 15,272; Control N = 20,475

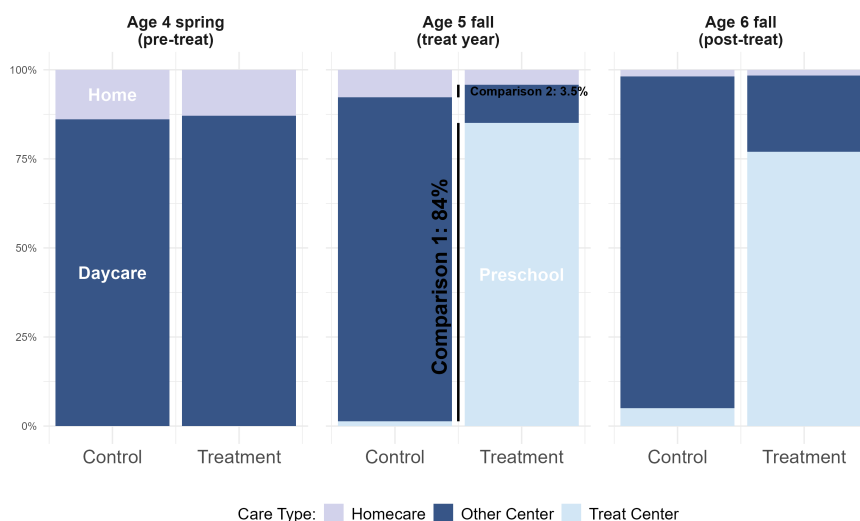


Figure 5 shows that before the experiment, centre-based ECEC was already common: in the spring preceding assignment, about 87% of children in our sample were enrolled in ECEC, while the remaining 13% were in home care.^[13] The experiment strongly shifted children into two-year preschool. In the autumn, 85% of children assigned to the treatment group enrolled in two-year preschool, compared with only 1.5% of children assigned to the control group. Correspondingly, regular ECEC participation at age five was much more common in the control group (92.3%) than in the treatment group (10.7%). Importantly, the experiment also moved some children from home care into centre-based provision: 4.2% of children in the treatment group did not participate in ECEC at age five, compared with 7.7% in the control group. By age six, participation in preschool was almost universal in both groups.

These patterns are informative about who was affected by the reform and how. Roughly 84% of children were *compliers* (comparison 1 in figure 5), meaning that their ECEC form was determined by the randomisation. Among these compliers, about 95% would otherwise have attended standard ECEC at age five, while only around 5% would have remained in home care. This distinction is central to interpreting the results: for the vast majority of children, two-year preschool replaced an already common and in many ways similar form of ECEC rather than a home environment.

Nevertheless, it is important to note that the experiment increased centre-based ECEC participation at age five by about 3.5 percentage points (comparison 2 in figure 5) from a baseline of roughly 92%. This increase was concentrated among children with Finnish-born parents; we find no corresponding rise in participation among children with immigrant backgrounds. The participation response was also stronger in the second year of implementation: assignment increased age-five participation by about two percentage points for the 2016 cohort and five percentage points for the 2017 cohort. Despite this increase, we find no effects on parental employment.

Taken together, these participation patterns highlight that the experiment changed childcare arrangements in different ways for different children. For most participants, two-year preschool replaced standard ECEC at age five, while for a smaller minority, it replaced home care. This distinction is crucial for interpreting the effects on children's skills: treatment effects depend not only on what the intervention provides, but also on what it replaces. In the next subsection, we therefore examine effects separately by counterfactual care environment, rather than relying solely on average treatment effects.

13. The estimates reported in Figure 5 differ slightly from those presented in the final evaluation report (Sarvimäki et al., 2026). The difference reflects a variation in sample construction: here we restrict the analysis to children who remained in the 148 municipalities participating in the experiment, whereas in the final report children who moved out of these municipalities were included in the "homecare or unknown" category. This alternative classification slightly affects reported participation rates, but has no impact on the estimated effects of the experiment.

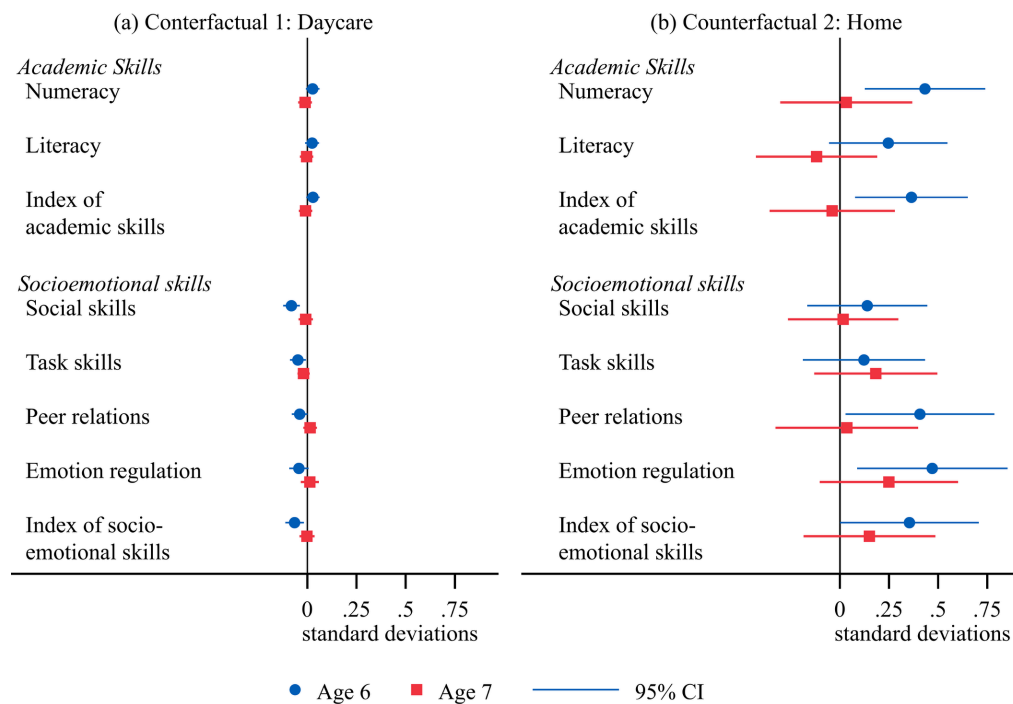
7.3 Counterfactual specific estimates

Figure 6 reports counterfactual-specific treatment effects for the two indices discussed above, as well as for the six preregistered subcomponents of academic and socioemotional skills.^[14] When the counterfactual is standard ECEC, the results are clear and closely mirror the intention-to-treat estimates discussed above. At age seven, the point estimate for the academic skills index is -0.01 standard deviations, with a 95% confidence interval ranging from -0.05 to 0.03 standard deviations. The corresponding point estimate for the socioemotional skills index is -0.002, and the confidence interval rules out effect sizes larger than 0.04 standard deviations in either direction. This similarity to the intention-to-treat results is expected, given the high compliance with treatment assignment and the fact that, for most children, two-year preschool replaced standard ECEC rather than home care. In short, replacing standard ECEC with two-year preschool did not alter the pace of children's skill development.

The results are very different for the small group of children whose alternative to two-year preschool would have been home care at age five. For this group, the point estimates at age six suggest sizeable and positive effects on both academic and socioemotional skills. The point estimate for the academic skills index is 0.36 standard deviations, corresponding to roughly half a year of typical developmental progress. The corresponding point estimate for the socioemotional skills index is very similar at 0.35 standard deviations, driven primarily by improvements in peer relations and emotion regulation.

14. Following Hull (2018) and Abdulkadiroğlu et al. (2016), these counterfactual-specific effects can be identified within an instrumental variables framework under mild additional assumptions beyond the standard IV setup using both comparisons from figure 5. We estimate the model using two-stage least squares with two endogenous variables, instrumented by treatment assignment and its interaction with ECEC enrolment status at the time of randomisation.

Figure 6. Treatment effects relative to daycare (Panel A) and home alternatives (Panel B)



By the start of primary education, however, these short-term differences appear to fade. At age seven, the point estimates are -0.04 and 0.15 standard deviations for the indices of academic and socioemotional skills, respectively. Importantly, however, statistical power is limited for this group because the number of children transitioning from home care to two-year preschool was relatively small, and the resulting estimates are imprecise. For the academic skills index, the 95% confidence interval ranges from -0.36 to 0.28 standard deviations, while for the socioemotional skills index, it ranges from -0.19 to 0.49 standard deviations among children whose counterfactual was home care. In short, the data do not allow firm conclusions about whether the sizeable initial gains for this group persist at school entry.

Lessons Learned: Counterfactuals Matter

Our results illustrate a general principle: the effect of an intervention depends critically on what it replaces. Two-year preschool produced large, short-term gains for children who would otherwise have been in home care, but essentially zero effects for children who would have attended standard daycare. Since 95% of the affected children fell into the latter category, the overall effect was close to zero. Future evaluations of early childhood programmes should anticipate this pattern and, where possible, design studies that can separately identify effects for different counterfactual margins.

8 Discussion

This paper discussed the practical implementation and the key results of Finland's nationwide randomised experiment on extending preschool from one to two years. The reform increased participation in formal early education at age five, moving some children from home care into centre-based provision. For these children, participating in two-year preschool had large positive short-term effects on both academic and socioemotional skills, although we cannot determine whether these gains persist into primary school. However, the potential of this channel is limited by the already high baseline enrolment rate. For most children, two-year preschool replaced standard daycare. For this large majority, extending preschool did not affect academic or socioemotional skills at school entry. Importantly, the scale of the experiment allows us to rule out even modest policy-relevant impacts for this group with confidence.

Beyond the specific question of preschool duration, this experiment provides a concrete example of how ambitious, system-wide reforms can be evaluated rigorously before permanent adoption. Randomised field experiments are not feasible or appropriate for every policy decision. However, this case illustrates that large-scale randomisation can be both feasible and informative even in complex policy environments such as early childhood education, in which implementation spans hundreds of centres and tens of thousands of children.

The aim of this paper was to support the design and implementation of similar efforts in the future. In our view, the main practical lessons from Finland's two-year preschool experiment are the following. First, the experiment required complementary expertise from multiple disciplines. Economists devised the research design and conducted most of the statistical analysis; educational scientists and psychologists developed the assessments, surveys, and documentation of pedagogical change; learning analytics experts implemented the large-scale testing infrastructure; and civil servants contributed essential legal, administrative, and institutional knowledge. This division of labour was critical for ensuring that the experiment was both policy-relevant and reliable.

Second, the experiment underscored the importance of close collaboration with civil servants and teachers. This co-operation went beyond securing formal consent or compliance. The trial succeeded because researchers and the civil servants jointly developed the design, and municipalities and teachers implemented the reform under real operational constraints. This shared ownership supported high participation and realistic implementation at hundreds of centres. Communication was itself a key input into design quality: clear explanations of randomisation to parents and staff, transparency about participation requirements, and tolerance for limited non-compliance when children's perceived interests or logistical constraints required it helped preserve trust while maintaining sufficient compliance for informative estimates.

Third, this trial illustrates what statistical power and careful design can buy. Many interventions are evaluated with samples too small to distinguish "no effect" from "the evidence is inconclusive", which leaves decision-makers free to project their prior beliefs onto noisy results. In contrast, the scale of this experiment delivered sufficiently tight confidence intervals to rule out effects that would be meaningful for the vast majority of children. The design also highlights the importance of targeting the right estimands: because treatment effects depend on what the intervention replaces, our counterfactual-specific estimates show that the relevant policy question is not whether preschool works in the abstract, but relative to which alternative and for which children. For countries considering similar reforms, the key implication is to design experiments that distinguish impacts relative to the main counterfactuals faced by the affected population.

We believe that these features have also been important for how the results have been received. Before the trial, extending preschool to two years enjoyed broad political support in Finland and featured in the campaign platforms of most major political parties. Parents and teachers likewise reported largely positive experiences with the two-year preschool during the experiment. Against this backdrop, the absence of improvements in children's skills relative to the existing system was probably an unwelcome surprise to many. Although the results have been public for only a month at the time of writing, early indications suggest that they are already shaping the policy debate. Importantly, the dominant response has not been to dismiss the findings, but rather to acknowledge the value of having conducted the experiment and of subjecting even popular and well-received reforms to rigorous evaluation.

Finally, we emphasise that large-scale experiments can strengthen data infrastructure even in countries like Finland, where researchers already have access to exceptionally rich administrative data. As a by-product of the trial, we developed detailed measures of children's skills and new information on educational inputs that had not previously been systematically recorded. The value of these data has been recognised by the current government, which has committed to continuing the assessments through

primary and lower secondary school. This will transform the experiment into the foundation of a new cohort study that can be maintained at relatively low marginal cost. As in other Nordic countries, the ability to link these new data to existing administrative registers also makes it possible to track children over the course of their lives. Such long-term research assets will inform not only the reform studied in this article, but a broad range of future educational and social policy questions.

References

- Abdulkadiroğlu, A., Angrist, J. D., Hull, P. D., & Pathak, P. A. (2016). Charters without lotteries: Testing takeovers in New Orleans and Boston. *American Economic Review*, 106(7), 1878–1920. Retrieved from: <https://doi.org/10.1257/aer.20150479>
- Antikainen, R. (2019). *Kokeilukulttuuri Suomessa – nykytilanne ja kehittämistarpeet* (Report 2/2019). Valtioneuvoston kanslia. Retrieved from: https://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/161281/2-2019-KOKSU_raportti.pdf
- Bronson, M. B., Goodson, B. D., Layzer, J. I., & Love, J. M. (1990). *Child behavior rating scale*. Abt Associates.
- De Wispelaere, J., Halmetoja, A., & Pulkka, V.-V. (2018). The rise (and fall) of the basic income experiment in Finland. *CESifo Forum*, 19, 15–19.
- Dunn, L. M., & Dunn, D. M. (2007). *Peabody Picture Vocabulary Test–Fourth Edition (PPVT-4)*. Pearson.
- Eerola, E., Kosonen, T., Kotakorpi, K., & Lyytikäinen, T. (2025). Tax compliance in the rental housing market: Evidence from a field experiment. *American Economic Journal: Economic Policy*.
- Einiö, E., & Nivala, A. (2025). *Rekrytointitukikokeilun loppuarviointi*. Retrieved from: https://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/166364/TEM_2025_28.pdf
- Goodman, R. (1997). The strengths and difficulties questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, 38, 581–586.
- Haaga, T., Kortelainen, M., Nokso-Koivisto, O., Saxell, T., Seppä, M., & Sääksvuori, L. (2025). Ostrobothnia digital clinic experiment. *ClinicalTrials.gov*. U.S. National Library of Medicine. Retrieved from: <https://clinicaltrials.gov/study/NCT06904469>
- Hämäläinen, K., Simanainen, M., & Verho, J. (2025). Health effects of cash transfers: Evidence from the Finnish basic income experiment. *Journal of Public Economics*, 250, 105480. Retrieved from: <https://doi.org/10.1016/j.jpubeco.2025.105480>
- Hämäläinen, K., & Uusitalo, R. (2005). Kannattaisi kokeilla: Kokeelliset menetelmät työvoimapolitiittisten toimenpiteiden vaikutusten arvioinnissa. Unpublished manuscript.
- Hämäläinen, K., Uusitalo, R., & Vuori, J. (2008). Varying biases in matching estimates: Evidence from two randomized job search training experiments. *Labour Economics*, 15(4), 604–618. Retrieved from: <https://doi.org/10.1016/j.labeco.2008.04.009>

- Hämäläinen, K., & Verho, J. (2022). Design and evaluation of the Finnish basic income experiment. *National Tax Journal*, 75(3), 573–596. Retrieved from: <https://doi.org/10.1086/720737>
- Harju, J., Kosonen, T., & Slemrod, J. (2020). Missing miles: Evasion responses to car taxes. *Journal of Public Economics*, 181, 104108. Retrieved from: <https://doi.org/10.1016/j.jpubeco.2019.104108>
- Hirvonen, S., Lassander, M., Sääksvuori, L., & Tukiainen, J. (2024). Mobilizing young voters with short text messages in nationwide field experiments. Retrieved from: <https://ace-economics.fi/kuvat/dp166.pdf>
- Holvio, A., Karhunen, H., Lerkkanen, M.-K., Nokso-Koivisto, O., Sarvimäki, M., & Torppa, M. (2024). Reading gains through teacher pairs: A co-teaching experiment. *Open Science Framework (OSF)*. Retrieved from: <https://osf.io/es3ng/overview>
- Hull, P. (2018). Isolating: Identifying counterfactual-specific treatment effects with cross-stratum comparisons. SSRN Working Paper. Retrieved from: <https://ssrn.com/abstract=2705108>
- Izadi, R., Luukkonen, E., Nokso-Koivisto, O., & Sarvimäki, M. (2020). Kaksivuotinen esiopetus -kokeilu: Tutkimussuunnitelma. Unpublished manuscript.
- Johansson, P., Rubin, D. B., & Schultzberg, M. (2021). On optimal rerandomization designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83(2), 395–403.
- Kangas, O., & Pulkka, V.-V. (2016). From idea to experiment: Report on universal basic income experiment in Finland. Kela Working Papers No. 106. The Social Insurance Institution of Finland (Kela). Retrieved from: <https://helda.helsinki.fi/server/api/core/bitstreams/7cf55de2-261e-481a-9242-adbf7f032a5f/content>.
- Karinen, R., Koivula, N., Koivula, T., et al. (2024). *Impacts and practical implementation of the Koto-SIB experiment*. Retrieved from: <https://julkaisut.valtioneuvosto.fi/handle/10024/166065>
- Kärnä, A., Voeten, M., Little, T. D., Poskiparta, E., Kaljonen, A., & Salmivalli, C. (2011). A large-scale evaluation of the KiVa anti-bullying program: Grades 4–6. *Child Development*, 82(1), 311–330. Retrieved from: <https://doi.org/10.1111/j.1467-8624.2010.01557.x>
- Kosonen, T., & Ropponen, O. (2015). The role of information in tax compliance: Evidence from a natural field experiment. *Economics Letters*, 129, 18–21. Retrieved from: <https://doi.org/10.1016/j.econlet.2015.01.026>

- Kraft, M. A. (2020). Interpreting effect sizes of education interventions. *Educational Researcher*, 49(4), 241–253.
- Lerkkanen, M.-K., Poikkeus, A.-M., & Ketonen, R. (2006). *ARMI: Luku- ja kirjoitustaidon arviointimateriaali 1. luokalle*. Helsinki: WSOY.
- List, J. A. (2022). *The voltage effect: How to make good ideas great and great ideas scale*. Crown Currency.
- Merrell, K. W. (1993). Using behavior rating scales to assess social skills and antisocial behavior in school settings: Development of the School Social Behavior Scales. *School Psychology Review*, 22(1), 115–133.
- Morgan, K. L., & Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2), 1263–1282. Retrieved from: <https://doi.org/10.1214/12-AOS1008>
- OECD. (2021). *Beyond academic learning: First results from the survey of social and emotional skills*. Paris: OECD.
- Onatsu-Arvilommi, T., & Nurmi, J. E. (2000). The role of task-avoidant and task-focused behaviors in the development of reading and mathematical skills during the first school year: A cross-lagged longitudinal study. *Journal of Educational Psychology*, 92, 478–491.
- Pekkala Kerr, S., Pekkarinen, T., Sarvimäki, M., & Uusitalo, R. (2020). Post-secondary education and information on labor market prospects: A randomized field experiment. *Labour Economics*, 66, 101888. Retrieved from: <https://doi.org/10.1016/j.labeco.2020.101888>
- Pesola, H., Sarvimäki, M., & Virkola, T. (2025). Randomization as an incentive device: Evidence from public procurement of immigrant integration services. Unpublished manuscript.
- Rege, M., Størksen, I., Solli, I. F., et al. (2024). The effects of a structured curriculum on preschool effectiveness: A field experiment. *Journal of Human Resources*, 59(2), 576–603.
- Salmivalli, C., & Poskiparta, E. (2012). KiVa anti-bullying program: Overview of evaluation studies based on a randomized controlled trial and national rollout in Finland. *International Journal of Conflict and Violence*, 6(2), 294–302. Retrieved from: <https://doi.org/10.4119/ijcv-2920>
- Sarvimäki, M., Holvio, A., Kuusiholma-Linnamäki, J., et al. (2026). *Kaksivuotisen esiopetuksen kokeilu: Loppuraportti*. Unpublished manuscript.

Tuulio-Henriksson, A., Simanainen, M., & Simanainen, M. (2020). Koettu terveys, psyykkinen hyvinvointi ja kognitiivinen toimintakyky. In O. Kangas, S. Jauhiainen, M. Simanainen, & M. Ylikännö (Eds.), *Suomen perustulokeilun arviointi* (Reports and Memorandums of the Ministry of Social Affairs and Health 2020:15). Ministry of Social Affairs and Health. Retrieved

from: https://julkaisut.valtioneuvosto.fi/bitstream/handle/10024/162219/STM_2020_15_rap.pdf

Verho, J., Hämäläinen, K., & Kanninen, O. (2022). Removing welfare traps: Employment responses in the Finnish basic income experiment. *American Economic Journal: Economic Policy*, 14(1), 501–522. Retrieved from: <https://doi.org/10.1257/pol.20200143>

Vuori, J., Silvonen, J., Vinokur, A. D., & Price, R. H. (2002). The Työhön job search program in Finland: Benefits for the unemployed with risk of depression or discouragement. *Journal of Occupational Health Psychology*, 7(1), 5–19. Retrieved from: <https://doi.org/10.1037/1076-8998.7.1.5>

Woodcock, R. W., & Johnson, M. B. (1977). *Woodcock-Johnson psycho-educational battery*. Riverside Publishing.

Comments on Ramin Izadi and Matti Sarvimäki: Learning to Experiment

Morten Visby Krægpøth

The article by Izadi and Sarvimäki (2026) provides an enlightening and relevant description of the implementation of a nationwide two-year preschool experiment. It takes readers through the stages of developing the experiment, from how the policymakers agreed to a large and complicated experiment to the process of randomly allocating treatment, to discussing the data collected. Unfortunately, the last two phases were not completed by the deadline for submitting these written comments, making it impossible to comment on the experiment's outcomes.

The experiment described looks at the effects of extending the existing one-year preschool to two years, meaning that children would enrol in preschool at age five rather than age six. Before the experiment, most five-year-olds were enrolled in daycare, meaning that the experiment, in some sense, tests the effects of enrolling five-year-olds in preschool rather than daycare.

It might be beyond the scope of the article to discuss in depth the differences between these two settings (daycare vs. preschool), but it is difficult for a non-Finn to grasp what is at stake here. One of the great strengths of the treatment description is, therefore, that the authors show how the quantifiable educational inputs have changed between the treatment and control groups.

I look forward to delving into the results when they are available. Researchers rarely have the opportunity to design experiments on this scale, and in such vital parts of the Nordic welfare model – childcare and schooling. Investing in early childhood education is considered highly cost-effective. At the peer-review conference, Sarvimäki made a strong case for greater knowledge of the effects of investments in early childhood education in settings where these investments are already relatively significant. I agree. To interpret the famous Heckman curve in the Nordic countries, we need to account for the baseline for additional investments. In this sense, this experiment is highly relevant.

On the other hand, I am slightly concerned – not having seen the results and the reactions to them – that the design, in a sense, puts all its eggs in one basket. By this, I mean that the article investigates the effects of a potential national policy to extend preschool to two years, but does not go into as much depth about the best way to implement such a policy change. I would expect the data to show local differences in

effectiveness, differences that might be the result of how successfully the change has been implemented. The extent to which the researchers have collected detailed data on local differences in adaptation is unclear to me.

Conducting such a large-scale experiment places significant responsibility on the researchers coordinating it and the civil servants with whom they work. For the experiment to maintain its legitimacy, it is important that there is also broad support for the implementation. Data collection, in particular, can be perceived as disruptive and time-consuming, taking time away from the professional's core task – education. I am therefore curious about how Izadi, Sarvimäki, and their collaborators have balanced the need to collect reliable assessment data on the one hand (carefully laid out in the paper) with the need not to disrupt more than necessary on the other. I would worry that if, for example, teachers find the data collection time-consuming and of no benefit to the children, the experiment might lose (some of) its legitimacy, and policy makers might be more reluctant to experiment in the future.

Comments on Ramin Izadi and Matti Sarvimäki: Learning to Experiment

Ingeborg Foldøy Solli

This is an important and interesting article about a large-scale field experiment in the Finnish ECEC sector. The article is well written, the field experiment was well executed, and I very much enjoyed reading about it. Implementation of the research design into the field requires a number of adaptations, and the authors guide the reader carefully through all stages of the implementation. They also generously share not only their reflections and considerations on how best to implement the intervention, but also their retrospective evaluation of the choices they made and thoughts on approaches that might have been better. Researchers conducting similar field experiments in the future will greatly benefit from these insights.

The authors find only negligible effects of the intervention on academic and socioemotional skills, results that deviate from existing empirical evidence. In particular, a similar intervention conducted in Norway finds treatment effects of .12 and .13 on a composite index of measured skills at the post-intervention and one-year follow-up surveys (Rege et al., 2024). A possible explanation for the difference is that the interventions changed practice more in the Norwegian experiment than in the Finnish one: While the existing Norwegian preschool curriculum is very nonspecific and unstructured, with a tradition of emphasising free play rather than didactic learning, it appears that the Finnish tradition has a more didactic approach. The Finnish intervention built directly on the existing framework, extending it to five-year-olds, while the Norwegian experiment introduced a new, structured and play-based curriculum. This could explain not only the lack of treatment effects on children whose counterfactual would be childcare, but also the large effects on children whose counterfactual would be home care.

Another notable difference between the Finnish and Norwegian experiments is the scale: While the Norwegian experiment was confined to a specific region and 71 childcare centres that volunteered to participate, the Finnish one covered 956 centres throughout the country. It is possible that Norwegian centres volunteered to take part because they were particularly interested in and motivated to implement this new type of curriculum, and the results may not be generalisable to the full population of centres. This illustrates the importance of large-scale interventions as a knowledge foundation for policymakers before nationwide reforms are implemented.

While it is crucial for the quality of implementation to allow municipalities and centres the necessary autonomy to take ownership of the intervention, this also implies less control over the implementation and a threat to internal validity. In particular, to ensure sufficient resources to implement the intervention, municipalities relocated their most experienced teachers from control to treated centres. Since this implies that also control centres were affected by the intervention, and some even closed down, this should not be considered an integral part of a bundled treatment. I believe this issue deserves more attention in the article.

It appears that baseline test scores were not controlled for in the analyses, possibly because children in home care were not assessed at baseline. It would be useful to have clarification on this and to understand the considerations done on this issue.

Finally, it was interesting to learn about the impressive “culture of experimentation” in the Finnish government and among Finnish policymakers, an example to follow for other countries.

References

Rege, M., I. Størksen, I. F. Solli, A. Kalil, M. M. McClelland, D. ten Braak, R. Lenes, S. Lunde, S. Breive, M. Carlsen, et al. (2024). The effects of a structured curriculum on preschool effectiveness: A field experiment. *Journal of Human Resources* 59(2), 576–603.



Nordic Economic Policy Review 2026

Evaluating Active Labour Market Policies Using Randomised Control Trials

Johan Vikström

Abstract

This paper discusses randomised controlled trials (RCTs) of labour market policies in the Nordic countries. Evaluating labour market policies is essential to ensure that public resources are allocated to effective interventions that improve employment outcomes, yet credible evaluation is challenging because programme participation is often selective. RCTs address this challenge by creating comparable treatment and control groups and are therefore widely regarded as the gold standard in policy evaluation. Using illustrative examples, the paper illustrates key issues related to motivating, designing, implementing, and analysing RCTs in a labour-market policy context. The examples cover interventions such as job-search assistance, monitoring, private service provision, and light-touch behavioural policies. The paper discusses ethical considerations, randomisation designs, implementation challenges, and analytical issues including take-up, outcome measurement, cost-effectiveness, and general equilibrium effects. It concludes that while the existing evidence base is substantial, continued experimentation and causal evidence remain important as labour market institutions and policies evolve. Building institutions that support RCTs is identified as one challenge.

1 Introduction

Evaluating labour market policies is essential to ensure that public resources are used effectively and that policies achieve their intended goals. Labour market interventions – such as training programmes, job-search assistance, or wage subsidies – are often costly and can have far-reaching effects on individuals, employers, and the broader economy. Without systematic evaluation, policymakers risk keeping ineffective or even harmful programmes going, while potentially overlooking initiatives that deliver substantial benefits. Rigorous evaluation provides credible evidence on what works, for whom, and under what conditions, enabling more efficient policy design and better targeting of support. In addition, evaluation can help uncover the mechanisms behind observed outcomes, offering insights into why certain programmes succeed or fail.

Conducting credible evaluations is challenging, however, because it is often difficult to find a credible comparison group. A non-experimental evaluation typically compares job prospects for unemployed job seekers who participate in a labour market programme or intervention with those of a comparison group consisting of unemployed job seekers who do not participate. The issue is that job seekers and caseworkers make choices based on individual circumstances, motivation, preferences, and constraints, which can lead to bias in comparisons between participants and non-participants. As a result, those who end up participating in a policy intervention may differ systematically from those who do not, and adjusting for these differences between the treatment and control groups may be difficult or even impossible. As a result, non-experimental evaluations struggle to distinguish between the programme's true effects and differences that simply reflect who participates in it.

Randomised controlled trials (RCTs) provide a powerful way to overcome these challenges. By randomly assigning individuals, firms, or regions to treatment and control groups, RCTs create comparable treatment and control groups, enabling credible causal inference. They eliminate selection bias due to self-selection from the policy intervention. In doing so, RCTs generate credible, actionable evidence that can guide better policy design, improve accountability, and ensure that public funds are spent on initiatives that make a real difference in people's employment prospects.

Over the years, several RCTs of labour market policies have been conducted in the Nordic countries. It started with a series of experiments carried out in Finland, Denmark, and Sweden. As early as 1996–97, an RCT of a week-long job-search training workshop was conducted in Finland. The experiment was inspired by a previous RCT conducted in the US (see Vinokur et al., 2000). The aim was to change the role of participants from passive unemployed individuals to active job seekers as a means of helping them find jobs again. The evaluation conducted by Vuori et al. (2002) reveals a positive effect on self-reported job stability but no effect on employment rates or earnings.

This was followed by RCTs conducted in Sweden in 2004 (Hägglund, 2011, 2014) and in Denmark in 2005–06 (Graversen and van Ours, 2008a). For the Swedish RCTs, Hägglund (2014) compares the relative efficiency of combining job-search monitoring and job-search assistance with monitoring alone, while Hägglund (2011) estimates the effects of an invitation to a meeting aimed at monitoring search activity and assisting with more effective job searches. The Danish RCT (called *Hurtig i gang*) was implemented in two counties and consisted of early and mandatory participation in a job-search assistance course, frequent meetings with employment officers, and programme participation after a few months of unemployment. An initial evaluation by Graversen and van Ours (2008a) found large positive effects on employment. Subsequent evaluations of the same experiment include Graversen and van Ours (2008b), Rosholm (2008), Blasco and Rosholm (2011), Vikström et al. (2013), Sørensen (2016, 2017) and Gautier et al. (2018).

Since those early experiments, a variety of RCTs have been conducted in the Nordic countries. Broadly speaking, Denmark was an early mover, with several experiments between 2005 and 2011 including the *Hurtig i gang* RCT mentioned above, the follow-up RCTs (*Hurtig i gang 2*) evaluated by Rosholm and Svarer (2009a), the youth intervention (*Unge – godt i gang*) evaluated by Høeberg et al. (2011), an intervention targeting long-term unemployed people (*Alle i gang*) evaluated by Rosholm and Svarer (2009b), as well as experiments on early meetings (Maibom et al., 2017) and outsourcing labour market policies to private-sector providers (Rehwald et al., 2017).

More recently, Sweden has been particularly active in conducting RCT evaluations of a wide range of topics, including Engström et al. (2012), Bennmarker et al. (2013), Laun et al. (2014), Helgesson et al. (2020), Fogelgren et al. (2023), Egebark et al. (2024), Dahlberg et al. (2024), Cheung et al. (2025), Hensvik et al. (2025), and Cockx et al. (2025).

RCTs in Norway include Bjorvatn et al. (2021), Berg et al. (2021), Sveinsdottir et al. (2021), and Hernæs (2025). In Finland, they include Malmberg-Heimonen (2005) and Pesola et al. (2025). Note that this is not an exhaustive list of all RCTs; in particular, it excludes some interventions targeted at individuals with mental health issues and interventions related to disability insurance that are less directly connected to labour market policy.

These RCTs span a range of different policies, including evaluations of early and frequent meetings with caseworkers and other job-search initiatives (Hägglund, 2014; Cheung et al., 2025), activations programmes (Graversen and van Ours, 2008ab; Rosholm, 2008; Vikström et al., 2013; Gautier et al., 2018), group and remote meetings (Maibom et al., 2017), private providers (Bennmarker et al., 2013; Laun et al., 2014; Egebark et al., 2024; Hernæs, 2025), online job recommendations (Hensvik et al., 2025), light-touch interventions with behavioural aspects (Bjorvatn et al., 2021; Cockx et al.,

2025), supported employment (Berg et al., 2021; Fogelgren et al., 2023), early intervention programmes for immigrants (Sveinsdottir et al., 2021; Dahlberg et al., 2024), other immigration integration services (Pesola et al., 2025), intensive job-search assistance for newly arrived immigrants (Joona and Nekby, 2012; Helgesson et al., 2020), vacancy referrals (Engström et al., 2012), intensified monitoring (Hägglund, 2014), displacement effects (Gautier et al., 2018; Cheung et al., 2025), information about rules and regulations (Cairo and Mahlstedt, 2023), policies that combine monitoring and job-search assistance (Hägglund, 2011) as well as interventions that combine job-search support and psychological aspects (e.g., Vuori et al., 2002).

At the same time, more evidence is needed to support efficient policy design. New policies are frequently introduced, and existing ones are often reformed, which may alter their effectiveness. More evidence is also needed on which programs work best for which groups. In some cases, it may be difficult to translate results from one country to another due to institutional and labour market differences. These are just some of the many reasons why additional RCTs remain valuable, despite the growing body of evidence from previous Nordic studies.

This paper does not provide a full review of the existing experimental and non-experimental evidence. Instead, it presents examples of conducted RCTs and uses these examples to illustrate key aspects of motivating, designing, implementing, and evaluating RCTs in a labour-market policy context. To this end, Section 2 introduces three main RCT examples referred to throughout the paper. These include a Danish RCT of private providers, a Swedish RCT of early meetings, and a Norwegian RCT featuring a light-touch intervention aimed at promoting good habits.

Section 3 then analyses topics related to motivating, conducting, and implementing RCTs. It discusses the main advantages of RCTs in a labour market context and explains why they are considered the gold standard for understanding which labour market policies work and for whom. This section also covers ethical considerations when conducting RCTs, different randomisation designs, pre-analysis plans, implementation challenges, and research independence. One conclusion is that RCTs are often ethical to conduct. When difficult allocation decisions must be made, and resources are limited, random assignment can be as ethical as any other method for assigning job seekers to an intervention. This section also touches on several practical aspects of implementing RCTs and concludes that careful implementation is key.

Section 4 focuses on topics related to analysing RCTs. Even though RCTs are designed to allow simple comparisons between randomised treatment and control groups, several important factors should be considered when conducting and analysing them. These include issues related to take-up and the importance of using objective outcomes. In addition, to using objective outcomes, it is also important to measure the costs of policy interventions, as cost-effectiveness is often a central concern when

assessing outcomes. Section 4 also illustrates how the Nordic countries are well endowed with excellent administrative register data and how this can be leveraged in RCTs. A complicating factor when conducting RCTs is the presence of general equilibrium (GE) effects – that is, when support for some job seekers may displace employment or affect non-treated job seekers in various ways. Such responses and their implications for RCTs are also discussed in Section 4.

2 Three examples of Nordic RCTs

Over the years, several RCTs on labour market policies have been conducted in the Nordic countries, most notably in Sweden and Denmark. This paper does not review all Nordic RCTs. Instead, it provides three examples on three separate research topics from three Nordic countries. These examples are then used to illustrate key aspects of RCTs in the subsequent sections.

2.1 Private providers in Denmark

Historically, employment services have been provided by government or municipal job-centers. In recent decades, however, many countries have moved toward and/or experimented with contracting out various employment services to private providers, e.g., training courses and job-search assistance, either giving private providers full responsibility for all employment services provided to the job seeker or contracting out only certain parts (e.g., a single training course).

To study the effects of private providers, Rehwald et al. (2017) evaluate a Danish RCT which compares private and public employment services. One underlying motivation is that, in theory, it is ambiguous whether public or private employment services are more effective. On one hand, competition among private providers and financial incentives may spur innovation (e.g., novel ways to help job seekers) and cost savings (e.g., less bureaucracy and greater resources spent helping the job seekers). On the other hand, contracting out involves transaction costs for setting up, monitoring, and regulating markets. It may also lead to unwanted cost cuts if private providers generate profits by lowering the quality of services (e.g., fewer meetings) provided to job seekers. Thus, can market provision deliver on its promises, or should employment services be provided in-house by public employment agencies?

The RCT evaluated by Rehwald et al. (2017) was conducted in 2011, with newly registered unemployed university graduates as the target group. The background is that, in Denmark, the municipalities are responsible for organising employment services, and in 2011, contracting out services to private providers was the default option for unemployed academics. That is, the municipalities were required to contract

out services. The RCT included 3,107 job seekers in four municipalities. The treatment group received the default option to go to a private provider, and the control group received support directly from the municipal job centres.

Private-provider systems may be organised in different ways, and their efficiency may depend on the exact design. In Denmark, 17 private providers were selected through a tendering round, where interested providers submitted offers, including a bid price for the total cost per job seeker per year. The public agencies then assigned job seekers to these selected providers, and the providers were not allowed to refuse any job seekers. The private providers were responsible for helping job seekers return to work using an employment-oriented strategy. Given basic requirements regarding the number of meetings and programme participation, the providers were free to organise the support as they saw fit. The compensation was heavily focused on a performance-related bonus (75% of the total compensation and 25% as basic compensation), awarded if the job seeker obtained regular employment for at least 13 weeks.

The findings from Rehwald et al. (2017) can be summarised as follows. On average, there are no differences in labour market outcomes (regular employment, subsidised employment, benefit receipt, or unemployment) between participants receiving public and private services. However, private provision is significantly more expensive, making it less cost-effective from a public spending perspective. Without transfer payments, the costs were DKK 10,392 for job seekers assigned to private services, while the costs for the public option were DKK 2,144, i.e., 21% lower. In terms of strategies, private providers held more frequent meetings and delivered more intensive, employment-oriented, and earlier interventions. By contrast, client satisfaction was reported to be slightly higher among jobseekers assigned to the public provider.

These conclusions align with results from several other RCTs of private provision. Focusing on RCTs from the Nordic countries, Bennmarker et al. (2013) analyse the effects of contracting out placement services for hard-to-place jobseekers in Sweden. They find that private providers generally do not improve labour market outcomes in terms of employment rates, earnings, and months worked. Unfortunately, they are not able to study costs. In a slightly different setting, Laun and Thoursie (2014) examine the effects of vocational rehabilitation for the long-term sick and temporarily disabled in Sweden and find no difference in labour-market outcomes or costs between private contractors and public services. More recently, also in Sweden, Egebark et al. (2024) evaluate a large-scale RCT of various aspects of private providers. One result is that the private providers were substantially more expensive (46-60% more, depending on the public sector support with which it is compared), but they did not generate more employment or higher labour earnings. They also show that private providers that are paid more do not generate better employment results.

Compared to the other studies, Rehwald et al. (2017) distinguish themselves in several ways. It provides the first Danish RCT evidence on contracting out employment services. It focuses on academics, which is relevant, as they may face a different labour market from other job seekers. It also evaluates a case involving the outsourcing of the entire package of employment services. The authors are able to leverage exceptionally rich data, such as reliable and frequent administrative data on employment, as well as detailed data on costs. The latter is key, as the merits of private providers largely hinge on the cost side, given that private providers do not better employment outcomes.

2.2 Early meetings in Sweden

Evidence suggests that early meetings and other types of job search assistance, aimed at helping job seekers search for jobs more efficiently, is one of the more powerful labour market policies (Card et al. 2010, 2018). This evidence includes results of several Nordic RCTs (see, e.g., Maibom et al., 2017 for a discussion). Early meetings may have positive direct effects on those treated. However, they may also lead to substantial displacement of jobs for the non-treated. In particular, if early meetings or any other labour market policy help some treated job seekers find jobs, this may come at the expense of other job seekers searching for jobs in the same local labour market. That is, ALMPs may lead to displacement effects, whereby jobs that non-treated individuals would otherwise have obtained are taken by treated participants. This is crucial from a policy perspective as displacement mainly represents a reordering of job queues rather than net employment gains. However, compared to the direct effects of early meetings, credible evidence on such displacement effects is scarce. Exceptions include Crépon et al. (2013), Gautier et al. (2018), and Ferracci et al. (2014).

This type of displacement of jobs represents one aspect of general equilibrium (GE) effects. Other GE responses may include changes in wage bargaining, matching efficiency, or firm behaviour—such as vacancy creation and market entry. For example, if an intervention helps job seekers search more efficiently, firms may fill vacancies faster, making it more profitable to post additional vacancies and thereby reducing unemployment. Conversely, if a policy strengthens workers' bargaining positions (e.g., by improving outside options), wages may rise, reducing firm profitability and potentially lowering employment. Such GE responses are highly relevant to consider when evaluating labour market policies.^[15] For instance, imagine a policy intervention that helps treated job seekers find jobs faster, leading to the conclusion that the policy

15. ALMPs may also create spillover effects on spouses, children, or peers. While a comprehensive evaluation ideally accounts for these channels, they are often considered second-order effects relative to the primary impacts on participants and the local labour market.

is effective. However, if this policy generates a high degree of job displacement or distorts wage bargaining, these negative GE employment responses could outweigh the positive employment effects for treated job seekers – ultimately reversing the conclusion about the policy's effectiveness.

Considering the importance of displacement effects and the scarce evidence, Cheung et al. (2025) implemented and evaluated a Swedish RCT specifically designed to capture displacement of jobs. The policy intervention consisted of three extra meetings with a caseworker during the first quarter of unemployment for all newly unemployed job seekers. To capture job displacement, these early meetings were evaluated using a two-level randomisation design, with randomisation over both job seekers within local offices and across local offices. The within-office randomisation of job seekers identifies whether the treated job seekers are more likely to find a job (the direct effect). This is similar to a standard RCT design with a randomised treatment and control group. The across-office randomisation identifies displacement and answers to what extent this comes at the expense of non-treated job seekers in the same labour market. Essentially, this part of the analysis compares non-treated job seekers at offices where half of the job seekers are randomised to the treatment (active offices) with non-treated job seekers at offices without any treated (non-active offices). Since neither group is called to early meetings, this difference measures the job-displacement effect arising from increased competition from treated job seekers in the same office.

The study documents a positive direct effect of early meetings on the re-employment rate by about 4.6%. However, it also finds substantial displacement, as the re-employment rate for the non-treated job seekers decreases by 3.8%. The implication is that failing to account for the job displacement would overstate the average impact by 89%. Nevertheless, the effect on overall employment, taking both direct and displacement effects into account, is still positive. The substantial displacement effects are consistent with previous studies (Crépon et al. 2013; Gautier et al. 2018). Face-to-face and distance meetings have equally positive employment effects, supporting the case for using communication technologies to provide job search assistance.

Cheung et al. (2025) also study the mechanisms behind these effects. They find that early meetings are effective because they increase the number of vacancy referrals passed from caseworkers to job seekers. Using their expertise, caseworkers are able to direct job seekers to relevant job openings, which helps job seekers apply to the most suitable positions at an earlier stage. By examining resource allocation, Cheung et al. (2025) also conclude that the observed displacement effects are due to job displacement rather than resources being reallocated from the control group to participants. Another result is that the effects of early meetings are more positive under favourable labour market conditions than in weaker labour markets with high unemployment rates.

Cheung et al. (2025) also shed more light on the GE effects. They analyse vacancy data and show that vacancies increase more in the active areas with the early meetings than in the control areas without any additional early meetings. The idea being that the improved job search as a result of the early meetings helps firms fill their open vacancies faster, making it more profitable to open more vacancies. Such a GE response contributes to the positive effects of the early meetings. Interestingly, Cheung et al. (2025) also find evidence of a delayed vacancy response, with larger changes in the vacancy rate toward the end of the experiment. That is, it takes time for firms to observe and react to the changes in job seeking caused by the early meetings. It suggests that job-search assistance may have more positive effects in the long term because, in the short term, the programme mainly leads to search congestion, while in the long term, improved searches induce firms to open more vacancies, increasing job finding rates and pushing down the unemployment rate.

To connect the different results, Cheung et al. (2025) combine the RCT variation with a structural analysis. The basic idea is to build a model that incorporates the improved job search, the job displacement, firms' creation of vacancies and any wage impacts. Without going into details, it involves using the estimates from the RCT to estimate a structural model and then use the estimated model for policy simulations. One simulation considers the employment effects of a full-scale rollout of the early meetings to all newly unemployed job seekers in Sweden. The simulations show that increasing the share of treated from 0% to 100% lowers the unemployment rate by around 0.2 percentage points, suggesting that the net effect of a full-scale rollout is positive despite the substantial displacement. The costs also decrease. When taking the observed delayed vacancy response into account and simulating a longer-term response to the early meetings, the overall assessment of early meetings becomes even more positive. This type of structural analysis has become more popular in recent years, and combining RCTs and structural analyses should be a key area for future research in the Nordic countries.

2.3 Light-touch intervention in Norway

The Norwegian RCT evaluated by Bjørvatn et al. (2021) begins with the idea that unemployment is a complex problem with many potential causes and no single solution. Traditionally, research has emphasised structural explanations such as inadequate skills, perverse incentives, and discrimination. However, a growing body of literature in psychology and behavioural economics suggests that behavioural factors, such as low motivation and bad habits, can lower job-finding rates among unemployed job seekers. There are several potential reasons for this. Unemployment may disrupt daily routines and lead to the loss of social connections, which in turn may foster bad

habits and negatively affect health and psychological well-being. Altogether, this may hinder an efficient job search.

Building on this background, Bjørvatn et al. (2021) designed and tested a light-touch goal-setting intervention aimed at counteracting harmful habits. The starting point was that the Norwegian public agency NAV identified a target population of 2,848 individuals aged 16-29 years in one province who had been unemployed for at least eight weeks. They received an SMS invitation to participate in a study, and the analysis sample consists of the 684 individuals who completed a baseline survey embedded in the experiment. It consisted of a set of background questions, and the goal-setting intervention evaluated in the RCT. This treatment was only given to the treatment group (two-thirds of the sample), and half of these participants also received SMS reminders about the goals. The remaining one-third in the control group only received the initial background questions in the survey.

Recognising that lifestyle improvements may play a crucial role in facilitating re-employment, the intervention focused on three "keystone" habits: sleep, exercise, and substance use. The underlying premise was that explicitly setting goals for healthier daily routines could enhance participants' prospects of finding employment. Specifically, it was divided into three steps: 1) questions about the job seekers' current situation; 2) asking participants to reflect on their current habits and express goals for the future; and 3) a summary of the expressed goals, along with encouragement to take a screenshot and remember them. The SMS reminder component of the intervention consisted of short weekly messages reminding participants of their goals. To evaluate the effects, the authors used both labour market outcomes based on administrative register data and follow-up surveys, focusing, for instance, on habits.

Fascinatingly, this very low-cost goal-setting intervention led to improved labour market outcomes. Twelve months after the intervention, the probability of being employed was 7 percentage points higher, and the share of participants receiving unemployment insurance benefits was 5 percentage points lower. This clearly shows that the intervention boosted employment rather than causing exits from the labour force. Moreover, about half of the participants reported wanting to exercise more and go to bed earlier, and these habits did improve following the intervention. Participants also demonstrated a more positive mindset and improved overall life satisfaction. There are also indications of an increased number of job applications. All this resulted from an essentially costless intervention without any apparent distortive effects on those already employed. The interpretation is that the act of setting goals itself may have generated a more optimistic outlook, reinforcing the changes in habits and contributing to improved labour market outcomes.

Bjørvatn et al. (2021) add to the growing literature on light-touch interventions. In terms of Nordic RCTs, one example is Cockx et al. (2025), who designed and evaluated an RCT of motivational emails designed to motivate job seekers to continue to actively

search for jobs. Hensvik et al. (2025) study how AI can alleviate search friction and help job seekers search for jobs online. Using an RCT, they evaluate an online job-recommender system that uses job seekers' click history to generate relevant personalised job recommendations. This study also relates to other non-Nordic RCTs. Two notable studies are Altmann et al. (2018) on the use of informational job-search folders and Belot et al. (2019) on the impact of low-cost online job-search advice.

3 Motivating, conducting and implementing RCTs

It is challenging to motivate, prepare and implement RCTs, and multiple pitfalls may ruin the trial. Some challenges relate to ethics, preparing pre-analysis plans, management of the trial, adjusting to unforeseen circumstances and questions related to research independence. This section discusses some implementation topics.

3.1 Motivating RCTs in the context of labour market policies

Designing effective labour market policies requires reliable evidence on what works, for whom, and under what conditions. Yet policy interventions are often complex, costly, and implemented in environments influenced by many confounding factors, making it difficult to identify their true impact. One key challenge is selection into a particular intervention, policy, or programme. A traditional policy evaluation based on non-experimental data compares the outcomes of treated participants with those of a comparison group, perhaps adjusting for observed background characteristics. This approach requires that the treatment and control groups are comparable in the absence of treatment. However, in practice, job seekers' and caseworkers' choices, actions, and preferences often mean that certain types of job seekers are more likely to be enrolled in certain policies. For instance, one policy may attract highly skilled and motivated job seekers, while another may attract those further removed from the labour market. Thus, without properly controlling for selection into the treatment group, comparisons between the treatment and control groups may lead to highly misleading conclusions, as such comparisons reflect both the selection and the policy and impacts. As a result, many non-experimental evaluations may capture selection effects rather than true causal impacts.

RCTs offer a powerful solution to these challenges. By randomly assigning individuals, firms, or regions to treatment and control groups, RCTs ensure that, on average, the treatment and control groups are comparable in terms of both observable and unobservable characteristics. This allows researchers to isolate the causal effect of an intervention, free from selection bias and other challenges (e.g., reverse causality). In

other words, the major advantage of RCTs is that they help resolve selection problems and provide credible answers to policy-relevant questions. To illustrate, consider the three RCTs above.

Rehwald et al. (2017) evaluate an RCT comparing public and private provision of ALMPs – a high-stakes policy question often shaped by ideological positions. Without randomisation, such evaluations typically rely on comparisons between job seekers who choose or are assigned to private providers and those who remain with public ones. Even after adjusting for observable characteristics, unobserved factors such as motivation or preferences could bias the results. If more motivated job seekers select private providers, the estimated effect would be positively biased; if less employable individuals are assigned to private providers, the bias would be negative. In either case, non-randomised comparisons risk producing misleading conclusions. This is shown explicitly by Bennis et al. (2013), who report that a non-experimental evaluation of private providers indicates a positive employment effect, whereas their RCT shows that there is no positive employment effect of private providers.

Cheung et al. (2025) evaluate frequent early meetings. In a non-randomised setting, job seekers called to more meetings may differ systematically from those called to fewer. Public employment agencies often allocate more meetings to individuals who need extra support, while stronger job seekers – who might find work quickly – receive fewer. As a result, simple comparisons could underestimate the real benefits of frequent meetings, since the control group would include stronger job seekers more likely to succeed regardless of support. Third, consider the novel intervention in Bjørvatn et al. (2021), which encourages job seekers to set personal goals related to daily habits such as sleep, exercise, and substance use. In a non-randomised implementation, agencies might target job seekers with poor baseline habits with such an intervention – a group likely to have weaker labour market attachment. This would then bias the estimated impact downward, underestimating the intervention's true effect.

In all these cases, explicit randomisation is essential to uncover the true causal impact and to avoid incorrect policy conclusions. While some non-experimental methods – such as exploiting rule changes, policy discontinuities, exogenous variation, or rich administrative data – can produce credible estimates, RCTs remain the gold standard, offering the most robust causal evidence. The importance of RCTs may differ, but generally, the benefits are greatest when selection into treatment is particularly challenging to address non-experimentally, there is limited prior credible evidence on the policy, and/or the policy is new, and random assignment is a fair and feasible allocation method.

Beyond estimating average effects, RCTs can also identify heterogeneous impacts, revealing which groups benefit most and informing policy refinement. RCTs also foster

policy learning and innovation, allowing decision-makers to test novel interventions – such as light-touch or behavioural policies – on a small scale before wider rollout. In short, RCTs bring scientific rigour to labour market policymaking, enabling governments to move beyond assumptions and ideology toward evidence-based interventions grounded in proper causal evaluation.

3.2 Ethics

When conducting RCTs, ethical considerations are important but rarely an obstacle. When considering RCTs to evaluate labour market policies, several key points emerge. First, public employment services regularly make complex decisions about assigning job seekers to programmes and interventions, aiming to provide them with the most suitable support. However, these decisions are often made in the face of significant uncertainty and without robust empirical evidence to guide them. Second, public employment services typically operate under binding resource constraints, meaning that not all job seekers can receive a given policy treatment.

Third, even though there is substantial research on active labour market policies, there are still many areas where evidence is lacking. For example, modern labour market policies may introduce new interventions that have not been previously evaluated. Evidence from one country may not necessarily generalise to other the Nordic countries. Policy designs may evolve over time, raising questions about their current effectiveness. Moreover, policymakers may wish to understand not only average effects but also heterogeneous effects across specific subgroups.

Taken together, these three factors provide strong justification for the use of RCTs. When tough allocation decisions must be made and resources are limited, random assignment can be as ethical as any other selection method. It also offers the additional benefit of generating credible causal evidence. This is particularly true when resource constraints mean that there is more job seekers deemed suitable for a given intervention than available slots. In other words, if there are explicit or implicit queues for a programme, determining access through random assignment may be just as ethical as attempting to make difficult prioritisations among job seekers. This is especially relevant when it is unclear whether a given programme helps or harms job seekers' chances of finding employment. In such cases, it may even be unethical not to conduct an RCT, since without experimental evidence, public agencies may continue implementing programmes that actually harm job seekers.

However, RCTs may become ethically problematic if these conditions are not met. For example, if there is strong prior evidence that a policy has positive effects and sufficient resources exist to treat everyone, then it would be unethical to randomise and withhold treatment from some job seekers. Similarly, legal or policy rules may

restrict randomisation. For instance, under Sweden's *Nystartsjobb* wage subsidy, employers are legally entitled to the subsidy if they hire a job seeker who has been unemployed for more than 12 months, which prevents randomisation.

Ultimately, these ethical considerations primarily concern the organisation implementing them (typically the public employment service), but researchers involved in RCTs must also carefully reflect on them. Formal ethical reviews should also be conducted if any sensitive personal data is processed.

3.3 Randomisation designs

A traditional RCT randomly assigns half of the units (e.g., job seekers) to a treatment group and the remaining half to a control group, which typically receives baseline services. Such RCTs estimate the effect of the evaluated treatment relative to the baseline, capturing the policy-relevant question of the additional direct impact of introducing the intervention. This standard design is a major improvement to non-experimental evaluations, but in some cases, it may also be valuable to consider more elaborate randomisation designs. One example is the displacement effects studied by Cheung et al. (2025), and the idea that early meetings that help some job seekers find jobs may displace jobs for other job seekers. Similar displacement effects may occur for other labour market policies, for instance, if training programmes or wage subsidies help treated job seekers secure jobs that would otherwise go to untreated individuals. To capture such displacement effects, randomisation at different levels, as in Cheung et al. (2025), may be important.

Two-level randomisation is one example of a more elaborate randomisation design. Other advanced designs include adaptive randomisation, which integrates random assignment with machine learning to improve allocation. Two-level randomisation can also be structured in various ways—not only across job seekers and local offices, but also across other relevant levels depending on the policy context. In cases with more than one type of treatment, such as in Bjørvatn et al. (2021), where a goal-setting intervention and a goal-setting intervention plus SMS reminders are the two treatments, it may be optimal, in terms of statistical precision, to randomise more than half of the job seekers to the treatment.

3.4 Pre-analysis plans

When designing and evaluating RCTs, a pre-analysis plan (PAP) that describes the programme/intervention, the experimental design, research hypotheses, as well as empirical models and outcomes, is important for several reasons. This is particularly true for RCTs when the researchers are involved in the design and implementation of

the intervention. One advantage of PAPs is that they require the researchers to clearly define hypotheses, outcome measures, and statistical methods before the experiment begins. By committing to these decisions in advance, PAPs reduce the risk of bias, selective reporting, and data mining. This strengthens the credibility, transparency, and replicability of experimental results, ensuring that findings reflect true effects rather than choices made after viewing the data.

In many fields, a pre-analysis plan is now also a requirement for publication in leading academic journals. However, some public policymakers may find PAPs restrictive, as they may wish to conduct exploratory analyses that address questions not specified in advance. While such flexibility is understandable, deviating from the PAP without documentation can undermine the credibility of the RCT, which is unfortunate since the central purpose of an RCT is to provide credible causal evidence on socially relevant questions.^[16]

3.5 Implementation challenges

Preparing and managing RCTs is often highly challenging, yet such challenges are rarely documented in research outputs. One exception is Cheung et al. (2020), which provides a dedicated implementation report for the RCT on early meetings evaluated by Cheung et al. (2025). This implementation report discusses some of the challenges.

First, as several other RCTs, Cheung et al. (2025) had an ambitious design intended to measure displacement effects, which required the mandatory participation of many local offices. Since participation was not voluntary in the first place, some offices lacked motivation, which made implementation more difficult. Opting for voluntary participation and randomising among local offices willing to take part in the experiment might have improved engagement and eased implementation. Similar issues can arise in other RCTs when the experimental protocol imposes specific requirements on caseworkers, local managers, or municipalities, potentially creating resistance or compliance issues. At the same time, only studying offices (or other types of participants) that enrol voluntary may create issues related to external validity if the volunteers are very different compared to other offices.

Second, many RCTs struggle to recruit enough participants. To ensure adequate sample sizes, the randomisation procedure must be applied consistently to the entire target group. If caseworkers are responsible for randomisation as part of their regular duties, the process must be simple, well-integrated, and clearly communicated. Automating randomisation as much as possible can also help. Likewise, effective tools to monitor the implementation of RCTs are crucial. In this respect, Cheung et al. (2020) conclude that future large-scale RCTs should use integrated data monitoring systems

16. Note also that PAPs can be revised when justified by significant new insights or circumstances—as long as all changes are clearly documented. Transparent revisions allow reviewers and other researchers to properly interpret the findings and maintain confidence in the study's integrity.

with real-time feedback and supportive supervision to ensure that the RCTs are fully implemented in accordance with the design of the experiment.

Third, successful implementation requires strong organisational anchoring and engagement from both operational staff and local offices. Cheung et al. (2020) highlight that the local offices involved in the trial received resources to conduct more frequent meetings but sustained local management support proved critical. Everyone involved needed to understand the purpose and benefits of the project. In particular, time invested in explaining the value of randomisation paid off by increasing acceptance and compliance with the experimental design.

These lessons underscore the importance of anchoring the RCT across the entire public employment agency implementing it and establishing a dedicated implementation team with sufficient resources to monitor all stages of the trial. Beyond ensuring compliance and proper delivery of the intervention, embedding serves other essential purposes. First, designing an effective RCT often requires detailed institutional knowledge, which only insiders within the employment agency can provide. Second, if the intervention is not aligned with existing structures, the results may fail to generalise, or the trial may encounter unforeseen barriers. Third, when an RCT is well-integrated into the agency's operations, the lessons learned are more likely to lead to sustainable policy improvements. Fourth, evaluating and interpreting the results often depends on access to detailed data and a deep understanding of the institutional context.

3.6 Building institutions that support RCTs

The overview of RCTs in the Nordic countries shows that Sweden has successfully conducted several RCTs over the past decade. This success is mostly due to the Swedish Public Employment Service having a dedicated group of highly motivated staff committed to generating policy-relevant causal evidence through RCTs, as well as to the close collaboration between this group and researchers at several academic institutions. This strategy has proven highly successful – without it, few, if any, RCTs on labour market policies in Sweden would have been implemented during the past decade. In particular, the dedicated team within the Swedish PES, together with its collaboration with researchers, has helped to motivate RCTs within the organisation, contributed to careful implementation, and ensured rigorous scientific design and the continued improvement of RCT-based evaluations. This has led to important evidence on policies such as early meetings with caseworkers (Cheung et al., 2025), private providers (Egebark et al., 2024), online job recommendations (Hensvik et al., 2025), supported employment (Fogelgren et al., 2023), early intervention programmes and job-search assistance initiatives for immigrants (Dahlberg et al., 2024; Helgesson et al., 2020), as well as light-touch interventions using email communication (Cockx et al., 2025).

The Swedish experience underscores the importance of building institutions that systematically support RCTs in the Nordic countries. Rather than relying on unique collaboration between a few highly motivated individuals within the PES and academia, the goal should be to establish durable institutional frameworks that facilitate rigorous experimentation and learning about an integral part of labour market policy design and evaluation. However, several challenges stand in the way of achieving this. One is that many actors may be reluctant to rely on evidence from RCTs, as such evidence can force politicians to reconsider ideologically motivated policies and compel bureaucrats within the Public Employment Service to abandon inefficient policies, which they have long supported and implemented. Some actors may also be so convinced of a policy's effectiveness that they view causal evidence from RCTs as unnecessary. In addition, as discussed above, implementing RCTs can be demanding, requiring dedicated resources for careful preparation and trial management.

Motivating and sustaining institutions capable of addressing these obstacles is itself challenging. Although Nordic contexts differ, building such institutions should be based on close collaboration between ministries, public employment agencies, and academic institutions. These organisations need to explicitly support RCTs and work jointly toward establishing durable institutional frameworks that facilitate experimentation and learning, without relying on a small number of dedicated individuals.

The Swedish experience also shows that close collaboration between public employment agencies and academic researchers can be an effective way to implement and learn from RCTs. Owing to this collaboration and the close involvement of researchers within the Swedish Public Employment Service at every stage of the RCT process, several evaluations have been co-authored by researchers from the Swedish PES and academic institutions (Cheung et al., 2025 is one example). This has received some criticism about the independence and credibility of the research output. However, several safeguards can mitigate such concerns: 1) A detailed pre-analysis plan, submitted before the analysis, ensures transparency and prevents undue influence being exerted on the evaluation; 2) Formal contracts between independent researchers and public employment agencies should specify the obligations of all parties and give the researchers the right to complete the evaluation autonomously if the public employment agency interferes with the evaluation. Moreover, implementing an RCT inherently requires significant involvement from public employment agencies, and any manipulation by them may occur during the implementation phase. Thus, what matters most for credibility is not excluding public agency staff from co-authorship, but ensuring transparency, independent analysis, and clear safeguards throughout the research process.

4 Analysing RCTs of labour market policies

Besides solving selection problems and providing credible evidence, RCTs also tend to simplify analysis, as they typically involve comparing the average outcomes of the randomised treatment and control groups. However, by boosting take-up, using objective outcome and cost data, exploiting high-quality Nordic administrative data, and carefully considering general equilibrium effects, additional policy lessons can be learned from RCTs. These aspects are discussed in this section.

4.1 Take up

In a standard RCT, take up refers to the difference in actual treatment participation between the randomised treatment and control groups. It depends on the share of individuals in the treatment group who ultimately receive the intervention and the share in the control group who receive it despite not being randomised to treatment. As such, it is closely related to the above-mentioned implementation challenges, and a high take-up rate usually indicates that the RCT was implemented according to the pre-analysis plan. At the same time, it is important to note that take-up rates well below 100% are quite common, but in most cases, this does not prevent meaningful analysis.

The RCTs mentioned in Section 2 all delivered robust and credible evidence on many policy-relevant questions, despite substantial differences in take up. For the early meetings experiment in Cheung et al. (2025), the take-up rate was 23%, meaning that 23% of the job seekers in the randomised treatment group were called in for extra meetings. This can be compared with a take-up rate of 100% in the light-touch habits intervention in Bjørvatn et al. (2021), as the analysis sample was defined as those who completed the online survey and the intervention. For the private-provider intervention in Rehwald et al. (2017), take up is slightly more complex, as the experiment ended prematurely in June 2012. This meant that the entire control group – who should have received public services but were eventually transferred to private providers – were also treated toward the end of the experiment. However, this does not invalidate the design, as the RCT could still measure the impact of private versus public support up to the end of June 2012. During this time, the difference in treatment exposure remained sizeable.

The take-up rates also vary across other Nordic RCTs, but this does not prevent meaningful analysis as long as there is a sufficiently large difference in take up between the treatment and control groups and if the sample size is large enough to detect a meaningful effect. For instance, a take-up rate of 20–30% can still yield valid results in a large sample. In such cases, there remains a clear difference in treatment exposure, and any observed differences in outcomes between the treatment and

control groups can still be credibly attributed to the intervention. In such cases, it is, however, instructive to also document and describe the various reasons behind any take up that is less than 100 percent. It helps to understand the implementation of the RCT and provides information about the average effect captured in the evaluation.

In these cases, traditional questions about which treatment effect to estimate become especially relevant. One option is to estimate the intention-to-treat (ITT) effect – the “reduced-form” impact of being assigned to treatment – measured as the simple difference in outcomes between the treatment and control groups. Since, most policies are not mandatory this may also be the most policy relevant effect to study.^[17]

At the same time, it is important to point out that high take-up is highly desirable, as it increases the statistical power of the experiment. When more job seekers in the treatment group actually receive the intervention, the evaluation captures more information about treatment outcomes, reducing uncertainty around the estimated average effects. By contrast, if there is no significant difference in take up between the treatment and control groups, the RCT has effectively failed, as there is no real variation in treatment exposure. In such cases, any observed differences in outcomes cannot be attributed to the policy being evaluated.

4.2 Objective outcomes

RCTs are designed to provide credible evidence on the effects of a policy. To accomplish this, it is crucial to use objective outcome measures that do not contaminate the RCT. The choice of outcomes depends on the data sources available, which may include self-reported assessments, caseworker reports, survey data, administrative records, or external sources such as tax data. Each type has strengths and weaknesses, but several general insights apply.

First, administrative data on unemployment spells collected by public employment agencies can pose challenges. These registers typically rely on caseworkers and job seekers accurately reporting when spells of unemployment end and why. Problems arise when incentives to report accurately are influenced by the intervention itself. For example, in evaluations of private providers, reporting responsibilities may shift to the providers, who could have incentives to record shorter unemployment spells if performance-based payments are tied to such outcomes. Indeed, in a recent evaluation of private providers in Sweden, Egebark et al. (2024) finds substantial differences between in-house unemployment data and more objective tax records that cannot

17. Alternatively, to capture the effect of actually receiving or participating in the intervention, an instrumental variables (IV) approach can be used. This method scales the ITT effect by the take-up rate, yielding the Local Average Treatment Effect (LATE)—the average effect for those who complied with their assignment and received the treatment. This distinction underscores why it is essential to study take-up behavior and to understand the characteristics of participants who ultimately engage with the intervention.

easily be manipulated. However, when the treatment does not affect reporting incentives, unemployment data from public agencies are more reliable.

Second, survey data can complement administrative data by capturing outcomes not observed in registers, such as job quality, health responses, or mechanisms underlying observed effects. Yet surveys face increasing challenges due to low response rates. Low participation reduces the effective sample size – for example, if only 10% respond – and may introduce bias if response likelihood differs by treatment status, undermining the RCT. Thus, while adding new data, surveys pose a challenge when used to construct key outcome data. Note also that in some settings, it is possible to achieve high response rates. One example is the habits intervention in Bjørvatn et al. (2021), which conducted follow-up surveys one and six months after the intervention and obtained response rates of 63% and 66%, respectively. One factor that probably contributed to the high response rates is that the intervention sample consisted of job seekers who had completed an initial online survey. Nevertheless, the high response rate is still impressive.

Third, some outcomes are conditional, meaning they are only observed for a subset of participants. In labour market studies, a common example is the wage of the first job after unemployment, which is only available for individuals who found jobs. If job-finding rates differ between the treatment and the control group, comparisons of wages may suffer from selection bias. This is a well-known issue with no simple solution, but it highlights the importance of carefully selecting and interpreting outcome measures.

4.3 Costs

Measuring outcome differences – such as the gap in employment rates between treated and control groups – is, of course, important. However, it is equally essential to consider the costs of an intervention. A small positive employment effect can lead to very different conclusions depending on whether the policy is cheap or expensive. Likewise, a negative employment effect may still be acceptable if the intervention is less costly than the baseline programme received by the control group.

That said, measuring costs is challenging, and several approaches can be used. From a strict cost–benefit perspective, all relevant costs (e.g., financial expenses, time, resources, or risks) and benefits (e.g., efficiency gains, improved outcomes, or social value) must be identified, assigned measurable – often monetary – values, and compared. In this framework, costs are all activities that consume resources, while benefits are those that create new resources. Thus, policies such as early meetings, training programmes, or rehabilitation measures count as costs, since they require staff time and related inputs. Conversely, wage subsidy programmes are not

considered costs in the strict sense, as they merely reallocate resources from public budgets to private firms or individuals (abstracting from administrative and supervisory expenses). In other cases, whether an expense is a cost depends on its economic nature. For example, reimbursements to private providers are costs when they finance services delivered to job seekers, but not when they generate profits, since strictly speaking, this represents a transfer from public to private actors rather than a net resource use.^[18] Similar complexities arise in measuring benefits. What is the economic or social value of reducing average unemployment duration by one month? The answer depends, among other things, on the production value associated with less time spent unemployed, which is sometimes difficult to quantify.

For these and other reasons, it is usually easier and in practice more common to measure impacts on public spending instead of a full-blown cost-benefit analysis. From such a public-spending perspective, the benefits of an intervention may come in the form of lower public expenditures (e.g., lower unemployment benefit payments) and public revenues (e.g., tax revenues). The costs are any increased public spending due to the policy of interest, including caseworker costs for early meetings, reimbursements to private providers, employment subsidies, etc. This public spending perspective differs from a strict cost-benefit analysis, as it focuses on impacts on budget spending rather than overall social welfare. However, a government spending analysis is still often policy-relevant, as the budgetary consequences of a policy are frequently of key importance to policy decisions. On a related note, often the focus is on budget spending (i.e., programme costs and unemployment insurance benefits), while ignoring budget revenue (e.g., tax revenues and social contributions).

In any case, when conducting RCTs, it is often important to take costs into account and to be explicit about the type of cost analysis (cost-benefit or public spending) being carried out. As already demonstrated, this distinction may have important implications for policy conclusions. To illustrate, consider the cost perspective in the three RCTs from Section 2. The habits intervention in Bjørvatn et al. (2021) was digital and self-guided, and thus almost costless to deliver once it had been designed. If scaled up, the cost per participant would be virtually zero, meaning that even small positive employment effects would be favourable from a cost-benefit perspective. However, there may still be limitations on the types and numbers of light-touch interventions job seekers can participate in.

Rehwald et al. (2017) conduct a detailed cost analysis focusing on the budgetary burden of private and public provision of employment services. The costs for the public programme can be divided into three components: 1) meeting costs in terms of caseworkers' time; 2) labour market programme costs; and 3) transfer payments. To

18. This holds if we abstract from any distributional considerations associated with redistributing resources across entities.

quantify meeting costs, they follow a previous study that estimated the time spent on each meeting and the hourly wage cost of caseworkers. Programme costs are calculated using the average cost per participant-week. Transfer payments include unemployment insurance benefits, social assistance benefits, and other individual-level benefits, as well as wage subsidies paid to firms. The costs of the private services follow a similar structure. They include direct payments made to private contractors, programme costs not borne by the private providers, and transfers, as in the public option. As already mentioned, the results of this cost analysis show that the private option is significantly more expensive while not producing better employment outcomes (partly because the analysis includes unemployment insurance and other benefits).

Finally, to evaluate the early-meetings program, Cheung et al. (2025) conduct a government spending analysis comparing reduced UI payments with programme costs. Benefit savings are estimated by multiplying the share of unemployed by their UI benefit levels under scenarios with and without the programme. Programme costs are the expenses for the caseworkers who conduct the early meetings. However, calculating these costs is challenging because the additional meeting expenses are not easily separable from regular caseworker costs at local offices; the accounting data only reports total caseworker costs. To overcome this, Cheung et al. (2025) use a time-use survey in which caseworkers reported how much time they spent on each meeting, including preparation, the meeting itself, and follow-up tasks such as documentation. Based on an average of 73 minutes of caseworker time per meeting, combined with information on caseworkers' monthly working hours, wages, and overhead, the authors calculate the cost per meeting and the total cost per participant for the early-meetings programme. This calculation demonstrates that it may be possible to obtain information about the costs of an intervention, even if the public employment agency does not explicitly separate programme costs from other costs. In terms of results, Cheung et al. (2025) find favourable government spending effects as the gains of the early meetings in terms of lower UI benefits outweigh the programme costs.

To summarise, the cost side is often key for policy but poses several challenges, for instance, related to separating the costs of a certain intervention from other costs incurred by public employment agencies. Hopefully, the examples above show that it is possible to conduct valuable cost analyses and that this should be an integral part of RCT evaluations, but this requires careful planning to assure that cost data is available for the evaluation.

4.4 Nordic data and analyses of mechanisms

Learning about employment effects and costs in a credible way is a major advantage of RCTs. In addition, several Nordic RCTs show that even more can be learned when

these experiments are combined with detailed Nordic administrative data. The Nordic countries have a distinct advantage in terms of individual-level data, which can further enhance what we learn from RCTs. For instance, using detailed data can help uncover the mechanisms behind observed effects – or the lack thereof – that is, help us understand why a certain policy worked or failed. This, in turn, can improve the design of future interventions and make it easier to translate lessons learned from RCTs to other contexts. Detailed Nordic data also enables the analysis of various additional outcomes, providing insights into relevant side effects of interventions, such as health consequences, the impact on families, and labour market exits, in addition to employment effects.

The researchers can achieve this in various ways. Take the early meetings evaluated by Cheung et al. (2025) as an example. They find that the early meetings boost job finding among the participants but also create displacement effects among non-participants in the same local market. The study also exploits the rich administrative data to study the mechanisms behind these effects. They exploit detailed administrative data from the Swedish PES on caseworkers (e.g., vacancy referrals and programme assignments) and job seekers (e.g., search behaviour). One result of this exercise is that early meetings are effective because they increase the number of vacancy referrals passed on from caseworkers to job seekers. Caseworkers use their expertise to find and point job seekers to relevant job openings. The jobseekers do not broaden their search in terms of occupations or geographical distance. Instead, the vacancy referrals lead to a more streamlined job search process by helping job seekers apply to the most relevant jobs earlier. The authors also find that face-to-face and distance meetings have equally positive employment effects, supporting the case for communication technologies when providing job search assistance.

This analysis uncovers the mechanisms behind the direct effects of early meetings for participants, but Cheung et al. (2025) also examine the mechanisms underlying displacement effects. Again, exploiting detailed Swedish data, they examine resource allocation and document that the negative effects for the non-treated do not arise because resources are allocated away from them – indicating that job displacement is the key mechanism. This finding is actually bad news for policy, as resource allocation issues can be mitigated by adjusting funding arrangements, whereas job displacement is far more difficult to address. In essence, the non-treated do not receive fewer referrals; rather, each vacancy receives more referrals, thereby increasing competition for available jobs and creating the observed displacement effects.

In summary, Cheung et al. (2025) use Swedish and Norwegian data, but similarly detailed and interesting administrative data are available in other Nordic countries. This is a great strength of labour research in the Nordic countries, a strength that is important to exploit when evaluating RCTs.

4.5 General equilibrium effects

The discussion of Cheung et al. (2025) and early meetings in Section 2.2. highlights the importance of general equilibrium (GE) effects. It includes job displacement, effects on wages and vacancies, and other more general responses to a given policy. Therefore, it is important to think about GE effects when designing RCTs. Key questions include: Is job displacement likely to be significant for the policy being studied? Could the intervention affect wages or the creation of vacancies? Might it affect how job seekers apply for jobs, and might such changes spill over to other job seekers? Several studies have also shown that accounting for such GE effects may be highly important for assessing the cost-effectiveness of labour market policies (Crépon et al., 2013; Gautier et al., 2018; Ferracci et al., 2014).

In addition to the more specific discussion of the general equilibrium effects of early meetings and other job-search assistance interventions in Section 2.2, several general points are worth noting. First, in some settings, GE effects may be extremely important to consider, as they can lead to very different policy conclusions, while in other settings, GE effects may be of second-order importance. Second, GE effects may either reinforce direct positive employment effects (e.g., the creation of new vacancies) or mitigate them (e.g., through job displacement). Third, in some contexts, it may be possible to capture GE effects within the randomisation design (e.g., a two-level design). Fourth, analysing GE responses in terms of vacancies and wages and incorporating them into structural analyses may enable more elaborate policy simulations and deeper policy learning beyond the original RCT.

5 Concluding remarks

This paper has used selected examples of randomised controlled trials from the Nordic countries to highlight central issues in motivating, designing, implementing, and evaluating various labour market policies using RCTs. Rather than aiming to provide an exhaustive review of the literature, the paper has illustrated how RCTs can be applied in practice and what can be learned from their use in real-world policy settings. One conclusion is that RCTs offer a powerful and often feasible approach for evaluating labour market policies. When resources are limited, and allocation decisions need to be made, random assignments can be ethically defensible and, in many cases, preferable to alternative allocation mechanisms. At the same time, the discussion in this paper emphasises that the credibility and usefulness of RCTs hinge on careful design and implementation, including transparent randomisation procedures, attention to ethical concerns, research independence, and practical constraints faced by the agencies implementing them.

Going forward, it is important to realise that conducting and evaluating RCTs can be challenging, and establishing institutions and collaborations that foster more RCTs and enable effective learning from them is important for all of the Nordic countries aiming to improve the efficiency of their labour market policies. Continued collaboration between researchers and policy institutions, combined with careful attention to design, implementation, and interpretation, will be essential for fully realising the potential of RCTs to inform effective labour market policy.

The paper also underscores that conducting and analysing RCTs involves more than simply comparing outcomes between treatment and control groups. Issues such as take-up, compliance, outcome measurement, and cost-effectiveness are central to interpreting results and informing policy decisions. The Nordic countries' rich administrative register data constitute a major advantage in this respect, enabling precise measurement of outcomes over time and across populations. However, the presence of general equilibrium effects, such as displacement of jobs, complicates interpretation, as positive effects for treated individuals may be partly offset by displacement or spillovers affecting non-treated job seekers.

Overall, the paper highlights both the promise and the limitations of RCTs in labour market policy evaluation. On the one hand, RCTs are generally regarded as the gold standard of evaluation methods, and in some cases, methods based on non-experimental data may lead to incorrect policy conclusions about the effectiveness of key policies. One example is research on private providers, where some non-experimental approaches suggest positive effects, whereas carefully conducted and reliable RCTs have shown that private providers are not a cost-effective way to apply labour market policies. This illustrates that RCTs are a crucial tool for generating credible evidence on what works, for whom, and under what conditions.

On the other hand, it is important that evidence from RCTs is complemented by evidence from other methodological approaches. In some cases, it is possible to exploit natural experiments to estimate causal effects in a credible way. For instance, if a policy is implemented only in certain areas, this may allow for credible comparisons across local areas, potentially also exploiting data from before the policy was introduced. In addition, several studies have exploited caseworker variation to obtain as-if random variation in policy assignment, effectively replicating an RCT without explicit randomisation (Cederlöf et al., 2025; Humlum et al., 2025). There may also be situations in which RCTs are not feasible, for example, due to ethical considerations, resource constraints, or institutional regulations. In such cases, non-experimental methods are often preferable to not evaluating.

References

- Altmann, S., Falk, A., Jäger, S., & Zimmermann, F. (2018). Learning about job search: A field experiment with job seekers in Germany. *Journal of Public Economics*, 164, 33–49.
- Berg, H., Hauge, K. E., Markussen, S., & Zhang, T. (2021). *Supported employment eller vanlig oppfølging? Resultater fra et stort randomisert forsøk i NAV*. Rapport 2021:2. Oslo: Frischsenteret.
- Bennmarker, H., Grönqvist, E., & Öckert, B. (2013). Effects of contracting out employment services: Evidence from a randomized experiment. *Journal of Public Economics*, 98, 68–84.
- Belot, M., Kircher, P., & Muller, P. (2019). Providing advice to jobseekers at low cost: An experimental study on online advice. *Review of Economic Studies*, 86, 1411–1447.
- Bjørvatn, K., Ekström, M., & Garcia Pires, A. J. (2021). Setting goals for keystone habits improves labor market prospects and life satisfaction for unemployed youth: Experimental evidence from Norway. *Journal of Economic Behavior & Organization*, 188, 1109–1123.
- Blasco, S., & Rosholm, M. (2011). The impact of active labour market policy on post-unemployment outcomes: Evidence from a social experiment in Denmark. *Mimeo*.
- Cairo, S., & Mahlstedt, R. (2023). The disparate effects of information provision: A field experiment on the work incentives of social welfare. *Journal of Public Economics*, 226, Article 104987.
- Card, D., Klüve, J., & Weber, A. (2010). Active labour market policy evaluations: A meta-analysis. *Economic Journal*, 120, F452–F477.
- Card, D., Klüve, J., & Weber, A. (2018). What works? A meta-analysis of recent active labor market program evaluations. *Journal of the European Economic Association*, 15(3), 894–931.
- Cederlöf, J., Söderström, M., & Vikström, J. (2025). The role of caseworkers: Job finding, job quality, and determinants of value-added. *Journal of the European Economic Association*. (Accepted).
- Cheung, M., Egebark, J., Forslund, A., Laun, L., Rödin, M., & Vikström, J. (2020). Implementation of a labor market program with more frequent meetings in Sweden. *IFAU Working Paper 2020:22*.
- Cheung, M., Egebark, J., Forslund, A., Laun, L., Rödin, M., & Vikström, J. (2025). Does job search assistance reduce unemployment? Evidence on displacement effects and mechanisms. *Journal of Labor Economics*, 43(1), 47–81.

- Crépon, B., Duflo, E., Gurgand, M., Rathelot, R., & Zamora, P. (2013). Do labor market policies have displacement effects? Evidence from a clustered randomized experiment. *The Quarterly Journal of Economics*, 128(2), 531–580.
- Cockx, B., Egebark, J., Van Hove, G., Videnord, E., & Vikström, J. (2025). Motivating job seekers: A field experiment. *IFAU Working Paper*.
- Dahlberg, M., Egebark, J., Vikman, U., & Özcan, G. (2024). Labor market integration of refugees: RCT evidence from an early intervention program in Sweden. *Journal of Economic Behavior & Organization*, 217, 614–630.
- Engström, P., Hesselius, P., & Holmlund, B. (2012). Vacancy referrals, job search, and the duration of unemployment: A randomized experiment. *Labour*, 26, 419–435.
- Egebark, J., Laun, L., Liljeberg, L., Rödin, M., Söderström, M., Videnord, E., & Vikström, J. (2024). *En effektutvärdering av arbetsförmedling med fristående leverantörer (An impact evaluation of employment services with independent providers)*. IFAU Report 2024:17.
- Ferracci, M., Jolivet, G., & van den Berg, G. J. (2014). Evidence of treatment spillovers within markets. *Review of Economics and Statistics*, 96(5), 812–823.
- Fogelgren, M., Ornstein, P., Rödin, M., Skogman Thoursie, P., & Thoursie, P. (2023). Is supported employment effective for young adults with disability pension? Evidence from a Swedish randomized evaluation. *Journal of Human Resources*, 58, 452–487.
- 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.
- Graversen, B., & van Ours, J. (2008a). Activating unemployed workers: Experimental evidence from Denmark. *Economic Letters*, 100, 308–310.
- Graversen, B., & van Ours, J. (2008b). How to help unemployed find jobs quickly: Experimental evidence from a mandatory activation programme. *Journal of Public Economics*, 92, 2020–2035.
- Helgesson, P., Jönsson, E., Ornstein, P., Rödin, M., & Westin, U. (2020). Equal entry: Can job search assistance increase employment for newly arrived immigrant women? *Arbetsförmedlingen analys*, 2020(10).
- Hensvik, L., Le Barbanchon, T., & Rathelot, R. (2025). Recommender systems and the labor market. *Mimeo, Uppsala University*.
- Hernæs, O. (2025). Performance pay for private program providers and impact on participants: A field experiment with employment services in Norway. *Labour Economics*, 95.

- Høeberg, L., Poulsen, J., Hertz, M., Svarer, M., & Rosholm, M. (2011). *Evaluering unge – god i gang*. Rambøll.
- Humlum, A., Munch, J. R., & Rasmussen, M. (2023). What works for the unemployed? Evidence from quasi-random caseworker assignments. *IZA Discussion Paper No. 16033*.
- Häggglund, P. (2014). Experimental evidence from active placement efforts among unemployed in Sweden. *Evaluation Review*, 38(3), 191–216.
- Häggglund, P. (2011). Are there pre-programme effects of Swedish active labour market policies? Evidence from three randomized experiments. *Economic Letters*, 112, 91–93.
- Laun, L., & Thoursie, P. (2014). Does privatisation of vocational rehabilitation improve labour market opportunities? Evidence from a field experiment in Sweden. *Journal of Health Economics*, 34, 59–72.
- Maibom, J., Rosholm, M., & Svarer, M. (2017). Experimental evidence on the effects of early meetings and activation. *Scandinavian Journal of Economics*, 119, 541–570.
- Malmberg-Heimonen, I., & Vuori, J. (2005). Financial incentives and job-search training: Methods to increase labour market integration in contemporary welfare states. *Social Policy & Administration*, 39, 247–259.
- Pesola, H., Sarvimäki, M., & Virkola, T. (2025). Randomization as an incentive device: Evidence from public procurement of immigrant integration services. *SSRN Working Paper*.
- Rehwald, K., Rosholm, M., & Svarer, M. (2017). Do public or private providers of employment services matter for employment? Evidence from a randomized experiment. *Labour Economics*, 45, 169–187.
- Rosholm, M. (2008). Experimental evidence on the nature of the Danish employment miracle. *Working Paper 08-14*. University of Aarhus.
- Rosholm, M., & Svarer, M. (2009a). *Kvantitativ evaluering af Hurtig i gang* 2. Arbejdsmarkedsstyrelsen.
- Rosholm, M., & Svarer, M. (2009b). *Kvantitativ evaluering af Alle i gang*. STAR Arbejdspapiret.
- Sørensen, K. L. (2016). Heterogeneous impacts on earnings from an early effort in labor market programs. *Labour Economics*, 41, 266–279.
- Sørensen, K. L. (2017). Active labour market programmes and reservation wages: It is a hazard. *Applied Economics Letters*, 24(9), 589–593.

Sveinsdottir, V., Evjen, T., Frangakis, M., & Opsahl, J. (2020). *Effektevaluering av Raskt i jobb for flyktninger: En randomisert kontrollert studie*. Rapport 2020:3. NORCE Norwegian Research Centre.

Vinokur, A. D., Schul, Y., Vuori, J., & Price, R. H. (2000). Two years after a job loss: Long-term impact of the JOBS program on reemployment and mental health. *Journal of Occupational Health Psychology*, 5(1), 32–47.

Vuori, J., Silvonen, J., Vinokur, A. D., & Price, R. H. (2002). The Työhön Job Search Program in Finland: Benefits for the unemployed with risk of depression or discouragement. *Journal of Occupational Health Psychology*, 7(1), 5–19.

Vikström, J., Rosholm, M., & Svarer, M. (2013). The effectiveness of active labor market policies: Evidence from a social experiment using non-parametric bounds. *Labour Economics*, 24, 58–67.

Comments on Johan Vikström: Evaluating Active Labour Market Policies Using Randomised Control Trials

Jukka Mattila

Johan Vikström has most definitely found an interesting subject – discussing the evaluation of labour market policy using randomised controlled trials (RCTs). The article explains why RCTs are useful for evaluating labour market policy, but its most interesting and arguably novel points concern the practical issues a researcher (or policy maker) faces in motivating, conducting, and implementing RCTs. Few, if any, economists contest the broad usefulness of RCTs, so it is perhaps not worth dwelling on this point for too long. The practical questions concerning experiments are more interesting, I think, also from a design point of view. Resource constraints and other practical limits mean that large-scale controlled experiments will probably always be a rare occurrence, but public officials, e.g., tax officers or employment services directors, often have sufficient operational leeway to conduct small-scale field experiments. As demonstrated by the highly fascinating Bjørvatn et al. (2021), an example cited by Vikström, such experiments might also yield important insights that could lead to significant improvements in the service concerned. As demonstrated by multiple examples, a culture of using RCTs to assess the impact of various programmes and actions, along with the researchers involved, would be a useful way of enhancing the effectiveness of the public sector. Why is it so difficult, then?

As Vikström notes, challenges in designing and implementing RCTs in practice are rarely documented in research outputs, even though there are many, ranging from ethical questions to design and effective implementation. Anyone who has attempted to facilitate or even managed an RCT is familiar with at least some of these issues. Experience in Finland has demonstrated that large-scale experiments are prone to various types of problems that can be difficult to foresee and sometimes to solve. Some of the issues faced by the Finnish basic income experiment are documented in Hämmäläinen & Verho (2022) and range from the reluctance of the tax authority to take part in the trial to other policy changes coinciding with the experiment. The Finnish two-year preschool experiment has been criticised for spillover effects between the control and treatment groups due to a shortage of trained early education teachers and the municipalities moving qualified teachers from the control group to the treatment group. Large-scale experiments are expensive investments, with the

budgets for both of the Finnish experiments just mentioned in the €20 million ballpark. It would be worth investing in avoiding mishaps in such cases. Some of the problems encountered might at least have been alleviated if shared experiences and good practices established by fellow researchers were more readily available. However, other problems persist.

Vikström lists many key subjects that deserve attention in the practical questions related to RCTs, and he provides a brief analysis of this broad topic. One of the more interesting points he makes is about institutions. In practice, it often seems that what really underlies an ethical argument against trials, for example, is not really about ethics, but about incentives – only wearing a disguise. Politicians might not have any motivation to explore whether ideological aims result in what they hoped (or claimed) they would – failure to achieve their goals might not be good for them. Civil servants dislike changing their routines and are reluctant to adopt new approaches. Vikström recounts the positive story of Sweden's public employment services dedicating in-house researchers to support RCTs, but the Finnish experience has been mixed. In unfortunate cases, in-house researchers in public agencies may vehemently oppose such experiments. They, too, might be uninterested in finding out whether the service they may have spent decades developing would stand up to the scrutiny of a rigorous controlled trial. Or, as Vikström also points out, they might simply be committed to the idea that what they are doing is so good that a trial is not warranted. They might lack the capacity to conduct such trials and be concerned for their status. Ultimately, they might have more to lose than to gain: the proposed services might turn out to be insignificant or, in the worst-case scenario, even to be harmful. In such cases, ignorance can be bliss. Institutions can then, instead of advocating for RCTs, curl up in a hedgehog-like position and simply oppose them by any means at their disposal. The first objection usually concerns ethics.

The core issue is that incentives matter. In its quest for identification, contemporary microeconomics appreciates RCTs for their value as evidence, so researchers have a natural incentive to be interested in them. They produce good publications. Other key stakeholders, politicians and civil servants, might feel less incentivised. In the worst cases, they may even have incentives not to participate or support such a policy, because it would require extra effort on top of their Monday morning chores, after all. The public sector is not incentivised by market competition in the same way as the private sector. Some big tech companies have a reputation for not only recruiting microeconomists but also for a habit of running a lot of A/B-testing, a close cousin to RCTs, when developing services. Unlike a tech company that faces market competition, the public employment service agency will probably still be there tomorrow, regardless of whether its work has any impact, and legislation decrees that it will have customers for its services. It would seem to me that the key question of incentives is the real problem that undermines our capability to develop a culture of RCTs in the field of not

only labour market policy but in the public sector in general – a culture that would probably serve us well by making public services more efficient and enhancing the wellbeing of the population.

It is also worth noting, as Vikström points out at the end, that even RCTs can have built-in problems and might not be the silver bullet people sometimes consider them to be. For example, controlled experiments (as opposed to natural ones) cannot always be blinded. As such, the participants – preschool teachers and the staff of public-sector employment agencies – may know that they are taking part in an experiment. Whether this affects their behaviour seems to be an open question, but placebo effects are well documented in other disciplines, such as medicine. If the simple act of participating in an interesting, possibly societally significant experiment that receives media attention is a motivating factor, it may raise concerns about external validity. To a policy maker or a civil servant, of course, the integrity of the internal validity of the experiment is somewhat less interesting if the external validity is compromised, because a lot of the value of an experiment lies in testing something with the interest of a full-scale rollout. A practical application of a policy can also probably never be a continuous randomised controlled experiment with a control group. If key incentives related to the experiment rely on for example a procurement contract that requires a control group (Pesola et al., 2025), it might be difficult to move from a trial to a full-scale service because the lack of a control group could ruin the incentive structure. This is not a criticism of Vikström. On the contrary, these points perhaps only underline the importance of the discussion Vikström initiates about developing better RCTs in practical environments. Carefully selecting the appropriate way to generate information is an important part of the researcher's understanding.

Vikström's article is an interesting and important starting point for a potentially fruitful and important discussion about RCTs. Society at large would benefit from an administrative culture that, as a matter of routine, seeks to improve the services it provides based on evidence from RCTs. Economics might also help us find ways to provide institutions with incentives to work together and to routinely design controlled trials alongside researchers as part of their normal work of improving services. The irony probably is, however, that a paper about institutions would perhaps have less value in the academic discipline right now than an experiment would, even though work on finding the ways to establish institutions that foster more of these kinds of trials would be crucial to conducting more of them. On this point, it is easy to agree with Vikström's concluding remarks that working on improving institutions and on co-operation between researchers and policy institutions (and the civil servants involved) would be important going forward, not just for labour market policy, of course, but for a great many policy spheres related to the provision of public services.

References

Bjørvatn, K., Ekström, M., & Garcia Pires, A. J. (2021). Setting goals for keystone habits improves labor market prospects and life satisfaction for unemployed youth: Experimental evidence from Norway. *Journal of Economic Behavior & Organization*, 188, 1109–1123.

Hämäläinen, K., & Verho, J. (2022). Design and evaluation of the Finnish basic income experiment. *National Tax Journal*, 75(3), 563–592.

Pesola, H., Sarvimäki, M., & Virkola, T. (2025). Randomization as an incentive device: Evidence from public procurement of immigrant integration services. *SSRN Working Paper*.

Comments on Johan Vikström: Evaluating Active Labour Market Policies Using Randomised Control Trials

Kari Hämäläinen

The credibility revolution in empirical economics has elevated the role of experimental evidence in recent decades. While excellent practical guides to the use of randomisation as a research design are available (e.g., Duflo et al., 2007; Rothstein and von Wachter, 2017), the article by Johan Vikström distinguishes itself by offering an insightful and valuable discussion of randomised controlled trials (RCTs) applied to labour market policies in the Nordic countries. Its focus is on the actual implementation of RCTs in a real-world policy context. Using three RCTs as examples, the article emphasises that conducting them requires substantial effort, careful planning, and strong institutional motivation, and it illustrates how researchers have addressed a range of challenges. It also highlights how the rich register data available in the Nordic countries can provide not only objective outcome measures but also tools to uncover the mechanisms underlying treatment effects and to derive additional policy lessons. Finally, the article underscores the importance of combining estimated treatment effects with programme costs to assess the feasibility of an intervention.

The article discusses several complications in designing RCTs that arise from operational, ethical, and administrative difficulties in randomly assigning study populations to treatment. These include imperfect compliance, deviations from the Stable Unit Treatment Value Assumption (SUTVA), endogenously observed outcomes, and multiple treatments. As a first comment, I note that one aspect that receives relatively limited attention is the distinction between the internal and external validity of RCTs. The appeal of RCTs lies in their ability to solve the assignment problem by creating comparable treatment and control groups. While this ensures high internal validity in well-conducted RCTs, concerns may arise regarding their external validity.

RCTs frequently rely on voluntary participation and/or tendering processes, raising questions about the extent to which their findings can be generalised beyond the study population. When members of the target population are invited to participate, those who choose to do so may differ in the distribution of the treatment effect from those who decline. A similar selection concern arises in tendering processes, as providers who expect to benefit from the intervention may be more likely to submit a tender. In such

cases, the effectiveness of a nationwide rollout may differ substantially from that observed in the original small-scale trial. For this reason, compulsory participation, where feasible, is strongly preferable from the perspectives of external validity and real-world effectiveness.

My second comment concerns the close, intensive collaboration required in conducting RCTs between researchers and the practitioners responsible for administering the intervention. A central insight of the article is that successful implementation depends critically on strong administrative support. Currently, when such support exists, it often relies on personal relationships between a small number of public sector employees and researchers. To improve this situation, the author calls for the development of institutional structures to support RCTs in the Nordic countries. I strongly back this proposal and would further argue for the adoption of formal quality assurance practices similar to those used in clinical trials. These typically involve independent oversight committees to safeguard data quality and trial integrity, as well as clearly defined responsibilities for the overall conduct and supervision of the intervention. The principal challenge of this proposal, however, is that it presupposes a genuine willingness to conduct rigorous assessments of the true effects of policies, and this commitment may be uneven or lacking in some parts of the public administration.

Another important issue addressed in the article concerns ethics, and the author correctly argues that RCTs are often ethically justified. In addition to the considerations discussed, I would like to add two further observations from the perspective of research ethics. First, power calculations are a necessary component of ethical study designs, as underpowered RCTs expose participants to interventions without a sufficient likelihood of producing informative results. When adequate power cannot be achieved, the study design should be revised, or the use of the RCT reconsidered. A further ethical consideration arises in the context of compulsory participation. Legislative changes frequently alter eligibility criteria for unemployment benefits and apply uniformly to affected populations, often without robust evidence of their impact. In such circumstances, trials conducted within a smaller population and with compulsory participation can be ethically justified, as credible estimates of policy effects can then be directly extrapolated to the broader population. Provided that appropriate protocols are in place, potentially harmful trials can be terminated early, and compensation mechanisms are more feasible for a limited group than for an entire population. Despite its compulsory nature, this approach is surely ethically preferable to the current practice of continuing potentially harmful policies that affect everybody who is unemployed.

Despite substantial progress in the evaluation of labour market policies, considerable gaps in our knowledge remain. The effects of specific interventions are likely to depend on broader labour market institutions, which vary significantly from country to country, and labour market policies themselves continue to evolve. These factors underscore the

ongoing need for rigorous empirical evidence, including well-designed RCTs, to assess policy effectiveness across different institutional contexts. In this regard, the article offers many valuable insights, and I would strongly recommend its inclusion on reading lists for ministries, employment agencies, and researchers concerned with the evaluation of labour market policies.

References

- Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomization in development economics research: A toolkit. In T. P. Schultz & J. A. Strauss (Eds.), *Handbook of development economics* (Vol. 4, pp. 3895–3962). Elsevier.
- Rothstein, J., & von Wachter, T. (2017). Social experiment in the labor market. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of field experiments* (Vol. 2, pp. 555–637). North-Holland.



Nordic Economic Policy Review 2026

Childhood and Adolescent Interventions in a Danish Setting: A Meta-Analysis of Evidence from TrygFonden's Centre for Child Research

Stine Nyhus Larsen, Nikolaj Noer Poulsen, Michael Rosholm and Katrine Bønneland Tølbøll

Abstract

This review and meta-analysis synthesises experimental and quasi-experimental intervention evaluations conducted at TrygFonden's Centre for Child Research (TFCCR) (updating and extending Rosholm et al., 2021). The objectives of the paper were to estimate pooled effects on children's learning and to explain between-study variation using cost-standardised effect sizes and coded information on intervention content, delivery and context, collected by means of a principal investigator survey. The eligible studies were RCTs or quasi-experiments that estimated causal effects on cognitive or educational outcomes, provided sufficient data to compute standardised and cost-standardised mean differences, and were conducted in Denmark by researchers affiliated with the Centre between 2013 and 2024. The articles were retrieved from [TFCCR's website](#). We applied a four-level random-effects model, assessed heterogeneity and small-study bias, and conducted meta-regressions aimed at identifying mediators and moderators. The meta-analysis encompassed 32 studies (49 treatment arms; 63 effect sizes), yielding a pooled learning effect size of 0.10 SD

(95% CI 0.07–0.14), which indicates positive but modest learning impacts, and a pooled cost-standardised effect of 0.04 per \$100 (95% CI 0.02–0.06). Effects tended to be larger for preschool-age targets, daycare settings, educator-focused delivery, and interventions that emphasise process quality. On average, the results did not differ between RCTs and quasi-experiments. The limitations of the study include short-run outcomes, potential cost measurement errors, clustering and spillovers, and generalisability, primarily to a Scandinavian welfare context. Consequently, the findings indicate that education interventions in a Danish setting produce small gains on average, with meaningful design- and context-driven variation in terms of both effects and cost-effectiveness.^[19]

1 Introduction

In recent years, there has been a growing interest in the effectiveness of interventions designed to improve educational outcomes. This research has mainly been motivated by the importance of education for income and wealth inequality, GDP, and growth (Barro, 2013; Hanushek & Woessmann, 2010), as well as by increasing inequality in income and wealth in many countries (Chancel & Piketty, 2021; OECD, 2024). In such situations, it has been suggested that improving the educational opportunities of disadvantaged children and youth will help to reduce inequality. For more than a decade, this line of research has been the focus of TrygFonden's Centre for Child Research at Aarhus University, Denmark.

TFCCR is an interdisciplinary research centre comprising Danish researchers primarily from political science, economics, psychology, and child language. The centre focuses on studying the effectiveness of interventions aimed at improving the learning and wellbeing of children and adolescents. It has accumulated a large body of evidence and knowledge, including how to support language development, both in general (Hoff et al., 2024; Højen et al., 2023) and for bilingual children (Andersen et al., 2021; Højen et al., 2021), the effect of teacher's aides in classrooms in a Danish context (Andersen et al., 2020, 2024), the effect of special camps for disadvantaged boys (Hvidman et al., 2020; Nielsen, 2023), the effects of interventions specifically targeting maths skills (Jensen & Sjø, 2024; Rosholm et al., 2025; Svane et al., 2023), evaluations of language tests (Bleses et al., 2022), and several other related topics.

Most of these evaluations have been conducted using randomised controlled trials (RCTs), which are considered the gold standard in causal inference (Imbens & Rubin, 2015a, 2015b; Schulz et al., 2010). In RCTs, identification depends on a relatively weak assumption of random assignment to treatment.

19. Note: This meta-analysis was not preregistered. The authors acknowledge funding from a grant from TrygFonden. TrygFonden was not involved in the production of this manuscript. The authors declare no competing interests. The data used in this study are available upon request from the authors.

However, RCTs are not always feasible or timely, and the availability of rich administrative register data in Denmark makes quasi-experimental methods a valuable complement for estimating the effects of educational interventions.^[20] Compared to RCTs, quasi-experimental methods are based on stronger and more fundamentally untestable assumptions, which are not solely reliant on statistical laws.

This study aims to summarise in a meta-analysis the accumulated body of research from both experimental and quasi-experimental studies conducted by TFCCR. One specific aim is to test the extent to which the use of experimental versus non-experimental methods influences average estimated effects. In addition, we investigate other potentially important moderators and mediators.

In the present paper, we update and extend previous work by Rosholm et al. (2021), who conducted a preliminary analysis on a set of interventions conducted by TrygFonden's Centre for Child Research, using cost-standardised effect sizes to document evidence for the Heckman curve (i.e., earlier investments yield larger returns) in a Danish context. We update the study pool by including additional RCTs and all available non-RCT causal impact studies. We also assemble cost data (where available) and systematise information on what the interventions consisted of, how they were delivered, and in which contexts they operated. As such, this current study contributes to existing literature on intervention efficiency in several dimensions:

1. The majority of previous childhood intervention studies have been conducted in the US. We present the in-depth results of interventions from a Scandinavian welfare state, which invests considerably larger resources in interventions aimed at children and youth (Rosholm et al., 2021).
2. All studies were conducted by researchers at the same research institution, and focused on children in the same country (Denmark), which provided methodological, cultural, and contextual homogeneity and consequently a high degree of comparability between studies.
3. Our study includes considerably more RCTs than the previous work by Rosholm et al. (2021), as well as all non-RCT causal impact studies conducted at the same research centre.
4. We take the cost side into account by only including studies for which we have access to information on intervention costs per participating child, which enables comparison between cost-standardised effect sizes.
5. We survey the principal investigators of the included studies to synthesise information on the content of the interventions, the characteristics of the deliveries, and the contexts, which allows for a detailed analysis of whether and how these potential features affect the cost-standardised effect sizes.

20. Quasi-experimental designs refer to designs that leverage naturally occurring variation to approximate random assignment, including difference-in-differences (DiD), regression discontinuity (RD), and instrumental variables (IV).

Note that while all of the studies adopt a methodological approach that identifies causal relationships, cross-study comparisons do not necessarily identify causal effects. Quasi-experimental studies are typically less generalisable than RCTs, and investigate effects among more narrowly defined populations. The differences in identification strategy mean that it is not feasible to directly compare causal implications between quasi-experimental studies and RCTs. Consequently, our study only asserts associations that may indicate interesting relationships between study characteristics and effect sizes.

The rest of the paper is structured as follows: Section 2 introduces existing literature on intervention studies. Section 3 presents the methods applied in the meta-analysis, including specifications of the four-level random effects model, and how heterogeneity and robustness of standard errors are assessed. It also presents specifications of the applied meta-regressions and assesses publication bias and influential papers. Section 4 presents the main results of the meta-analysis and meta-regressions. Section 5 discusses the results in the light of work in the field in general, and interprets the reported effect sizes. Section 6 addresses some of the limitations of the present study and how methodological choices affect the interpretability of the results. Section 7 presents conclusions.

2 Existing literature

Rosholm et al. (2021) surveyed in some detail the existing knowledge of the impact of early childhood education (ECE) programmes on children's learning, school interventions, and parent and family programmes. Here, we briefly summarise and extend their review.

In the United States, the impacts of ECE programmes have been widely researched via evaluations of programmes such as Head Start, pre-kindergarten, and other centre-based programmes, which have demonstrated significant short- and long-term effects (Elango et al., 2016). This research can be divided into two main lines: The first examines how access to broadly focused ECE programmes affects children's learning opportunities; while the second examines how the quality of children's preschool experiences can be enriched by evaluating the effects of specific instructional practices carried out in ECE settings (Byrne & Fielding-Barnsley, 1991; Clements & Sarama, 2007; Justice et al., 2010). In both lines of research, studies have consistently found that positive outcomes are associated with the provision of higher quality interactions between teachers and children, appropriately targeted curricula, enriched learning environments and teachers' professional development (Burchinal et al., 2010; Duncan & Magnuson, 2013; Fryer, 2017; Gardner et al., 2019).

Extensive research has been conducted on the factors that influence student achievement. Recent syntheses draw on more than 2,100 meta-analyses and over 130,000 individual studies (Hattie, 2023). However, there are relatively few studies with strong causal identification strategies, particularly in the context of school-based interventions with academic achievement as an outcome. The research within this area can be divided into two broad lines: The first focuses on structural factors, such as class size and school size; while the second studies what happens within the classroom in terms of teachers' instruction (Hanushek et al., 2003; Kane & Staiger, 2012). In general, studies tend to find positive effects associated with structural factors such as smaller class size (Fredriksson et al., 2013; Jepsen & Rivkin, 2009) and smaller school size (Humlum & Smith, 2015). Within the research on early reading instruction, there is also considerable evidence of the positive effects of systematic and explicit phonics instruction, compared to programmes that utilise primarily whole-word or whole-language approaches (Ehri et al., 2001; P. et al., 2000).

Research on parent and family programmes has not grown to the same extent as research on ECE programmes or school factors. There have been only limited strong causal evaluations of parent and family programmes, and there is a need for more research in this area (White et al., 2015). Despite the limited research, studies show that parent and family programmes can have positive effects on children's academic achievement, socioemotional development and long-term outcomes (Gardner et al., 2019; Lundahl et al., 2008; Reynolds et al., 2007). One example of a successful parent and family programme is the Abecedarian Project, which provided high-quality early care and education, parenting support, and health care to low-income children and their families, and has been shown to have long-lasting positive effects on children's development (Campbell et al., 2002).

In summary, the research on ECE programmes, school interventions and parent and family programmes has revealed wide-ranging evidence of overall positive impacts. However, the literature is not very precise with regard to the types of interventions that are most effective, and more research is needed to determine the most effective interventions for different groups of children (Levitt, 2009; Phillips et al., 2017).

3 Methods

3.1 Data

To identify relevant studies, we reviewed the complete list of published literature associated with TFCCR from its opening in 2013 until May 2024.^[21]

We limited ourselves to studies from this research centre, as it ensures consistency with respect to the type of data collected. Most importantly, it enables us to consistently construct cost-standardised effect measures for all studies, and to survey the principal investigators with regard to active ingredients, etc., which would otherwise be practically infeasible. It also means that all of the effect sizes originate in the Danish welfare state context. The effect size calculations utilise high-quality individual-level register data from the same databases, and the effect sizes are based on a consistent approach to intervention development. This creates homogeneity in some dimensions, which ultimately makes it easier to search for possible moderators in the relatively small sample. The main drawbacks of this inclusion criterion are that it means omitting insights from a much larger sample of potentially relevant studies, e.g., across the Nordic countries, and it carries a risk of self-bias. To mitigate the latter, we conduct a risk-of-bias analysis.

3.1.1 Inclusion Criteria

The above considerations give rise to the following formal inclusion criteria:

- The included studies were carried out in Denmark by a researcher affiliated with TFCCR.
- The included studies have concluded, and short-term causal impacts on cognitive/educational outcomes are available and published in either a scientific journal or a publicly available report.
- The included studies either report or allow for the calculation of standardised effect sizes and intervention costs by treatment arms (i.e., the additional costs of the intervention, net of evaluation costs, compared to the treatment as usual administered to the control group).

The inclusion criteria follow those specified in Rosholm et al. (2021), except that we also include studies evaluated using non-RCT methods. We include more studies than Rosholm et al. (2021) for two reasons: a) considerably more studies have since been completed, which means that more RCT results are available; b) we include studies

21. <https://childresearch.au.dk/udgivelser>

using non-RCT methods for causal analysis. In total, we include 32 studies, comprising 49 treatment arms and 63 reported effect sizes. These are summarised in section A in the supplementary material and in more detail in [online SI](#).

The authors mutually agreed on both the inclusion of studies and the collection of data points. We focused on identifying causal (unbiased) studies to minimise self-bias, as the research could otherwise be considered somewhat self-congratulatory. To further mitigate this, we include a risk-of-bias analysis, which was conducted using the risk-of-bias tools for randomised trials (RoB2) and non-randomised studies of interventions (ROBINS-I) developed by Cochrane methods (Sterne et al., 2019). One author assessed risk of bias for all studies, and each treatment arm and outcome measure was assessed individually.

3.1.2 Outcomes

For all studies, the effect size is calculated based on the effect of a standardised test score or an indicator of some educational/cognitive outcome. These outcome categories include the results of vocabulary, language and math tests, and indicators for obtaining a certain level of education. For more details, see [online SI](#).

For school-based interventions, learning outcomes are constructed using primarily register-based data for compulsory national test scores, which ensures a high degree of comparability across different interventions.

At present, there are no register-based outcome measures of cognitive skills for children in preschool-based interventions in Denmark. As such, the daycare teachers administered the language assessments, either through a published standardised assessment test, by employing a standard language test conducted by the municipality at the end of preschool, or by completing a standardised checklist. The study by Haghish et al. (2021) shows acceptable congruence between educators. Consequently, we do not believe that this has introduced any biases in the judgement of the children's skill level.

For interventions targeting young adults, the outcomes are typically indicators of obtaining a certain level of education (as the targeted groups were individuals without qualifying education).

When a single primary effect is reported on an outcome category, we use that as our indicator. Otherwise, if multiple effects are reported on subcategories within that outcome, we calculate the average effect size within that category. We use the formula for the variance of a sum of random variables to aggregate the corresponding standard errors. If an estimate of the correlation between the random variables is

available, we use it. Otherwise, we impute a conservative value of 1, which inflates the variance and assigns the lowest weights to that average effect. Generally, when individual standard errors are available, we use them directly. If there are no errors, we calculate them based on the test statistic or, in rare cases, request them from the principal investigator. The effect sizes are standardised by dividing the estimated effect by the standard deviation in the control group (Glass' Δ).^[22]

For comparison between interventions, it is relevant to consider not only standardised effect sizes, but also the associated costs. This entails gathering information on the cost of the intervention, the number of participants in the treatment and control group and the timing of intervention, which allows us to construct a cost-standardised effect size measure. We calculate the costs by measuring the extra costs of the intervention (net of research and evaluation costs, which include costs of data collection) over and above the costs of the usual treatment. We adjust the costs in DKK to 2021 levels, according to the Net Price Index,^[23] and convert the values to USD using the purchasing power parity (PPP) rate of 6.59 DKK per 1 USD. We assume a linear relation between costs and the number of participants. This means that we calculate the average cost of an intervention (or treatment arm) per participant, and divide the effect size and standard error by this cost to get an effect size per \$100. Studies for which it is not possible to retrieve a meaningful cost measure are excluded from the cost-standardised effect analysis. The reported learning outcomes and costs for all studies are reported in Table A. For more details, see [online SI](#).

3.1.3 Moderators and mediators

Meta-analysis has emerged as an important tool to synthesise findings from multiple studies, and summarise multiple findings. However, the overall effect size obtained from a meta-analysis can only provide limited insights into the underlying mechanisms that drive the results. A meta-regression analysis can help to identify the mediators (active ingredients) or moderators that explain the variability in effect sizes across studies.

To conduct the mediator-moderator analysis, for each study, we collect information on the average age of the treatment group, whether the effect was evaluated using an RCT (1/0), whether the intervention took place in preschool (1/0), and the type of learning outcome (vocabulary, language, maths, or other).

In addition to these potential moderators, we are also interested in the interventions' mediators or active ingredients. Hence, we surveyed principal investigators on the content of the interventions, as well as other complementary aspects. Specifically, we include information on:

22. If only the pooled standard deviation is available, we instead calculate Cohen's d . We prefer Glass' Δ because the treatment effect might not only affect the mean, but also the variance of the treatment group.

23. <https://www.statistikbanken.dk/pris116>

1. Who designed the intervention (categories: Researchers, Practitioners, Administrators/Policy makers/Others, NA)
2. Whether the intervention was aimed directly at a cognitive outcome or mediator, e.g., socio-emotional skills (categories: Directly at cognitive, Mediator, NA). We do not include a "both" category, and these cases are therefore modelled as 1 in both the cognitive and mediator columns.
3. Whether the primary focus of the intervention was on quantity, e.g., more lectures, or quality, e.g., lecture contents (categories: Quality, Quantity, NA). "Both" and "neither" are modelled as 1 in both columns and 0 in both columns, respectively.
4. Active ingredients (categories: Structural quality, Process quality, Learning format, Therapeutical elements, NA)
5. Whether the intervention was manualised (categories: Fully, To a medium extent, Not at all, NA)
6. Whether the intervention was targeted at a disadvantaged group (categories: Low cognitive skills or learning difficulties, Immigrant homes, Low SES homes, No, Other, NA),
7. The group targeted in the intervention (categories: Children, Parents, Educators, Other)
8. The location of the intervention (categories: Home, Camp, Daycare, School or educational institution, NA)
9. Whether and what type of support was provided during the intervention (categories: Course, Ongoing support, Written information, NA).

For details about individual studies, see [online SI](#).

3.2 Model specification

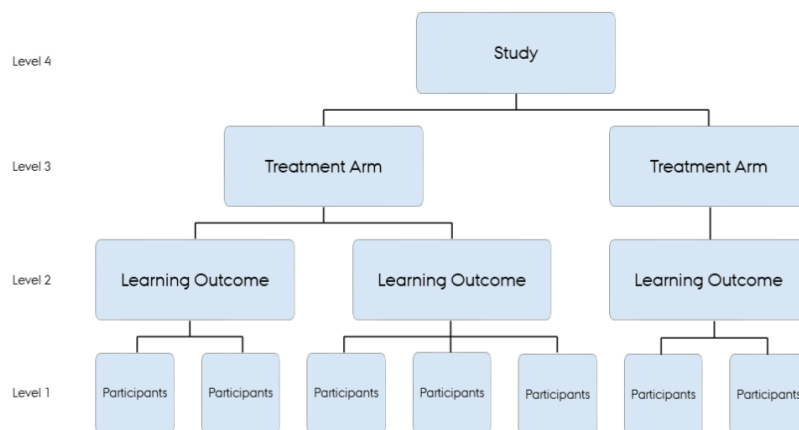
Our model specification considers the following two dynamics. First, the heterogeneity observed in our sample of studies may arise from the following three factors: i) the effect sizes reflect either vocabulary, language, or maths skill development; ii) the treatment is targeted to different subgroups of children aged 0–18; and iii) the content and implementation of the treatment vary significantly across studies within the same

target age. Second, we account for the interdependencies in our sample, as the same control group is used multiple times in comparisons across different treatment arms within the same studies, and the same treatment and control groups are often used to assess different learning outcomes. Consequently, we assume that the data-generation process has the following nested structure (see Figure 1): Dependency between different studies, as they are conducted at the same centre (level 4); dependency between the different treatment arms within each study (level 3); dependency between the different learning outcomes within treatment arms (level 2); and finally, dependency within the different learning outcomes, i.e., effect sizes, caused by participants (level 1).

Based on the nested structure, we employ a four-level random-effect model (for a mathematical specification, see section B.1). The model estimates the true effect size of a given outcome in a given treatment arm within a given study based on the true population effect size, plus a random component for each of the levels described above.

We test for heterogeneity of the effect size using Cochran's Q (Cochran, 1954) and quantify it using I^2 , which is independent of sample size and therefore easier to interpret. We follow the widely accepted guidelines, such that an I^2 value of 0 indicates no heterogeneity, while values of 25, 50 and 75 correspond to low, moderate and high heterogeneity, respectively (Higgins & Thompson, 2002). Formal specifications of heterogeneity and variance estimates are found in section B.2.

Figure 1. Visualisation of the hierarchy of the four included levels.



Despite the hierarchical structure of the model, some effect sizes within each cluster are based on the same control group. For that reason, we impose a cluster-robust inference method (also known as robust variance estimation). This method constructs

a cluster-robust sandwich estimate of the variance-covariance matrix of the model coefficients. This procedure follows the recommended general workflow for meta-analysis involving complex dependency structures (Viechtbauer, 2010).

We assume a correlation between language/reading and maths of 0.48, based on the average relationship magnitude found in Danish National Tests. Rho is used to model the correlation between sampling errors of two different outcome measures. This means that if two different learning outcomes are used to evaluate the same treatment arm, we assume a correlation between them of 0.48.

3.2.1 Meta-regressions

To test the association between outcomes and various potential moderators and mediators, we implement a mixed-effect meta-regression model. The model adds a vector of covariates containing the included mediators and moderators to the specification of the meta-analysis, but is otherwise the same. The mathematical notation is found in section B.3. The meta-regression model contains many variables. First, we remove those that are exact linear combinations (e.g., therapeutic elements as an active ingredient, which has a value of zero, as it is not used in any of the included studies). We then use a forward selection approach, based on Akaike's Information Criterion (AIC), to select variables for inclusion in the final models.

We estimate the parameters in the various models using R (version 4.2.3) (R Core Team, 2023), the metafor package (version 3.8.1) (Viechtbauer, 2010) and the meta package (Balduzzi et al., 2019).

4 Results

4.1 Meta-analysis

We now present the results of the two meta-analyses. First, we present the results of the estimates of the raw effect sizes; and second, the results of the estimates of the cost-standardised effect sizes.

4.1.1 Effect sizes

In this subsection, we present comparable results from all studies and treatment arms. Table 1 below presents the results of the meta-analysis based on the raw standardised effect sizes. We report coefficients and cluster-robust standard errors from the sandwich estimation.

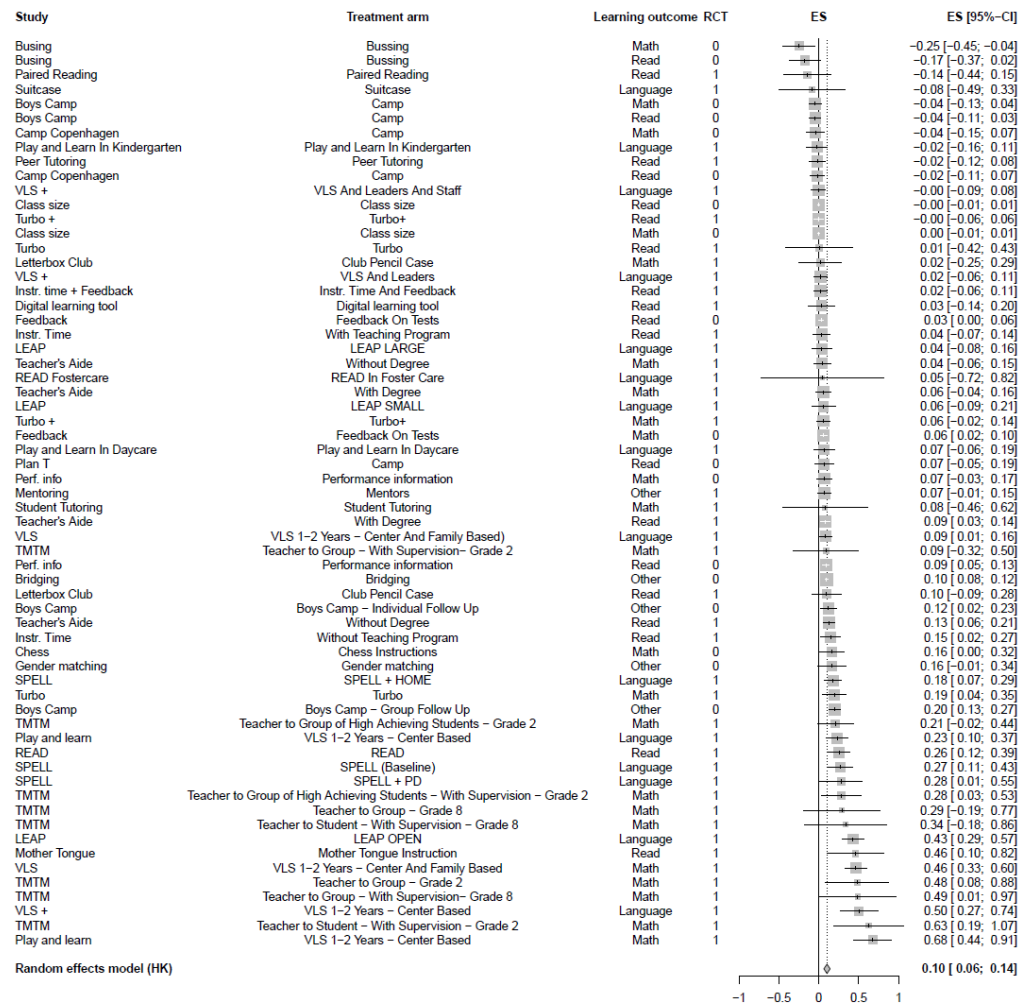
Table 1. Multivariate Meta-Analysis Results

Model Results					
Estimate	SE	t-value	p-value	CI (lower)	CI (upper)
0.104	0.020	5.265	< 0.001	0.065	0.144
Variance Components					
Component	Estimate	Sqrt	N levels	Fixed	Factor
σ_4^2	0.003	0.054	32	No	Study
σ_3^2	0.004	0.065	49	No	Study/Treatment arm
σ_2^2	0.008	0.088	63	No	Study/Treatment arm/ Learning outcome
Test for Heterogeneity					
Test	Df	Q-value	p-value		
Cochran's Q	62	370.009	< 0.001		

The estimate of the true population effect size, μ , based on the four-level meta-analysis, is 0.104 (95% CI: 0.065 - 0.144) and is statistically significant ($P < 0.001$). The three variance components (σ^2 's) show that of the total variation net of sampling error variance, 20% can be attributed to between-study heterogeneity, and 27% attributed to between-treatment-arms- within-study heterogeneity, while the remaining 53% can be attributed to within-treatment-arm heterogeneity. The I^2 statistic estimating heterogeneity was overall 95.0%, which suggests high heterogeneity. Calculating I^2 by hierarchical level revealed that 18.9% of heterogeneity was explained by variation at level 2 (within treatment arms), 26.6% at level 3 (between arms within studies) and 49.6% at level 4 (between studies).

Figure 2 depicts a forest plot of all included effect sizes for the different treatment arms and studies. The analysis includes a total of $i = 63$ effect sizes within 49 treatment arms across 32 studies. The observed effect sizes ranged from -0.25 to 0.68, with 79% of the estimates being positive, 52% of those being significant. The most effective intervention is Play and Learn (taking place in nurseries) with an effect size of 0.68, followed by one of the TMTM treatment arms (individual maths training in grade 2 in public schools), with an effect size of 0.63. For details of the studies, see section 5 and [online SI](#).

Figure 2. Forest plots of a random effects meta-analysis for different measures of effect size for the 32 studies, their 49 treatment arms and their 63 reported effect sizes. Whiskers on effect sizes represent confidence intervals, while the grey box represents weight as given by variance (more precise studies are given higher weight).



4.1.2 Effect sizes per \$100

In Figure 3, we plot effect sizes against the average cost of the intervention per participant, split by type of intervention (RCT or observational) and by statistical significance. The figure shows that, in general, there is no strong association between effect size and costs. It also shows considerable variation in costs per participant. For example, the Mentoring intervention (aimed at youth on social assistance) cost \$3,884 per participant, while the SPELL intervention (in kindergartens) cost \$92 per participant.

If we zoom in on the studies conducted using randomised trials (marked by round dots), it becomes apparent that these tend to have lower costs than those evaluated using other methods. However, whether we look at randomised trials or other methods, there is no obvious strong relation between the costs of the intervention and its effectiveness. This is somewhat surprising – based on cost-effectiveness considerations, we would have hoped for a positive relationship, such that more expensive interventions were also more effective. This motivates a follow-up calculation of the meta-analysis, utilising cost-standardised effect sizes. The aims are to determine how this changes the ranking of interventions in the forest plot and to enable an investigation of whether other characteristics may lead to differences in cost-standardised effect sizes. In our calculation, the cost-standardised effect size is defined as the effect size per \$100 per participant. The resulting estimates do not represent “real” effect sizes in the conventional sense, and the pooled effect size should therefore be interpreted with caution. However, as already stated, this approach makes it possible to measure the relative efficiency of the different interventions.

Not all of the studies involved a treatment associated with a measurable cost estimate (i.e. Class size). These studies were therefore excluded from the cost-effectiveness analysis. The analysis included a total of $k = 61$ effect sizes within 48 treatment arms across 31 studies. Table 2 shows the result of the meta-analysis of cost-standardised effect sizes. Again, we report cluster-robust coefficients and errors from a sandwich estimation. The pooled effect size, based on the four-level multivariate meta-analytic model and assuming correlation between learning outcomes, was $ES = 0.037$ (95% CI: 0.018 - 0.057).

The I^2 statistic was 99.9% overall, which again suggests a high level of heterogeneity. The level-wise calculation showed that 73.6%, 0% and 26.3% were explained by variation within treatment arms, between arms within studies and between studies, respectively.

Figure 4 presents a forest plot of the results across all studies, treatment arms and learning outcomes. The observed effect sizes ranged from -0.04 to 1.24, 80% of the estimates were positive, and 53% of those were significant. The most cost-effective treatments are information treatments (Performance info and Feedback). These take place in public-sector schools and do not have very large raw effect sizes, but have very low costs. The third most effective intervention is LEAP-OPEN (which takes place in kindergarten), which has a standardised effect size of 0.43 per \$100 invested. For more details on the most effective studies, see section 5 and [online SI](#).

Figure 3. Relationship between average cost per participant (2021 prices) and effect size

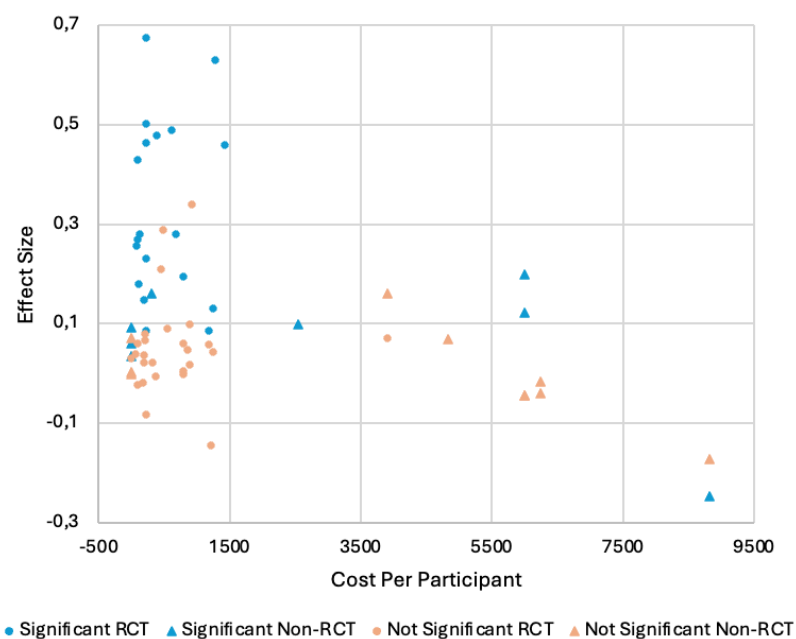
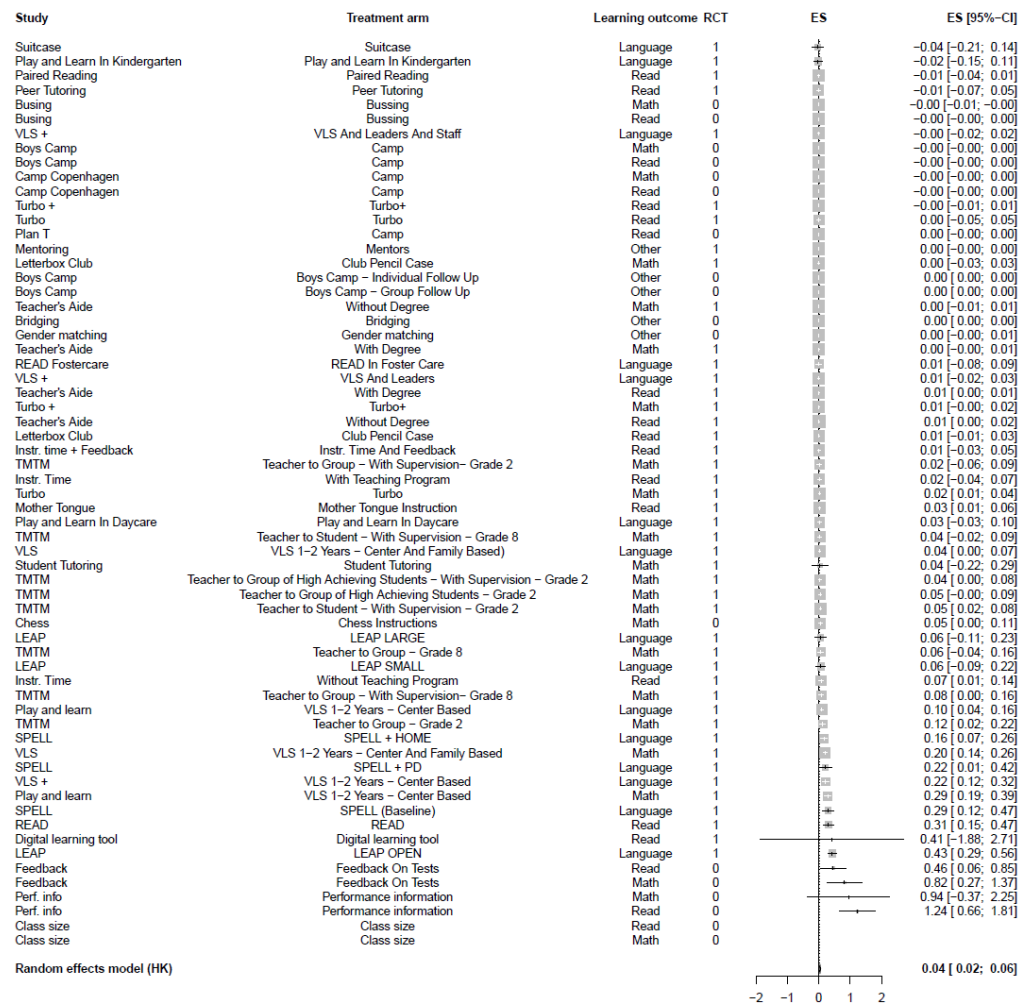


Table 2. Multivariate meta-analysis results: Cost-standardised effect size

Model Results					
Estimate	SE	t-value	p-value	CI (lower)	CI (upper)
0.037	0.010	3.826	< 0.001	0.018	0.057
Variance Components					
Component	Estimate	Sqrt	N levels	Fixed	Factor
σ_4^2	0.002	0.046	31	No	Study
σ_3^2	0.000	0.000	48	No	Study/Treatment arm
σ_2^2	0.001	0.027	61	No	Study/Treatment arm/ Learning outcome
Test for Heterogeneity					
Test	Df)	Q-value	p-value		
Cochran's Q	60	369.340	< 0.001		

Figure 4. Forest plots of a random effects meta-analysis for different measures of cost-standardised effect sizes for the 32 studies, their 49 treatment arms and their 63 reported effect sizes. Whiskers on effect sizes represent confidence intervals, while the grey box represents weight, as given by variance (more precise studies are given higher weight)



4.2 Meta-regressions

We conduct meta-regressions to investigate potentially important mediators and moderators in our model. We apply forward selection to statistically identify the mediators and moderators that best explain the variation in effect sizes. We code factor variables for various intervention and study-design choices presented as binary in section 3, and incorporate a continuous age measure. To avoid convergence problems and increase precision, we identify variables that are linear combinations in the mediator/moderator matrix. Variables identified as linear combinations of other variables are excluded prior to the forward selection procedure. We then choose the best model based on the AIC. Using BIC gives similar results, and the best-performing models are generally close in terms of information criteria. Table 3 presents an overview of the dummy variables included, their distribution of 0's and 1's, along with the variables selected in the procedure described below.

Table 4 presents the best model for the raw effect sizes, as determined by the forward selection procedure.

Table 3. Variables included in forward selection. Age is the only non-binary variable.

The mean age is 9.8 years. The column % 1 shows the percentage of interventions that include the mediator/moderator in question. The last two columns mark variables selected by forward selection (FS) for raw effect sizes and cost-standardised effect sizes, respectively.

	Variable	% 1	FS raw	FS cost
General design	RCT	71.4		
	Targeted preschool	27.0	*	*
	Age	-	*	
Learning outcome	Language	23.8	*	*
	Math	36.5		
	Other	7.9	*	
Active ingredients	Learning formats	65.1		
	Process quality	71.4	*	
	Structural quality	54.0		
	No data	9.5		
Aimed at cognitive outcome	Directly	87.3		
	Through mediator	36.5		
Intervention designer	Practitioner	71.4		
	Researcher	79.4		
	Administrator/others	65.1		
	No data	4.8		
Extent of manualization	Fully	12.7	*	*
	Medium extent	61.9		
	Not at all	19.0		
Placement of intervention	At camp	11.1		
	At daycare	17.5		*
	At school or institution	49.2		
	In everydaylife	1.6		
	Other	6.3		
Increasing quality or quantity	Quality	65.1	*	
	Quantity	57.1		
	No data	17.5		
Intervention support	Ongoing	23.8		
Group targeted at intervention	Children	66.7		
	Parents	6.3		
	Educators (pedagogues/teachers)	41.3	*	*
	Others	4.8		
Targeted disadvantaged group	Immigrant homes	6.3		
	Low skills/learning difficulties	38.1		
	Other	9.5		
	Not targeted	49.2		

Table 4. Multivariate meta-regression results: Raw effect

Variance Components					
Component	Estimate	Sqrt	N levels	Fixed	Factor
σ_4^2	0.000	0.000	32	No	Study.id
σ_3^2	0.006	0.080	49	No	Study.id/ treat.id
σ_2^2	0.000	0.000	63	No	Study.id/ treat.id/es.id
Test for Heterogeneity					
Test	Df	Q-value	p-value		
Cochran's Q	54	139.081	< 0.001		
Model Results					
	Estimate (SE)	z-value	p-value	CI (lower)	CI (upper)
Intercept	0.143 (0.067)	2.137	0.033	0.012	0.274
Targeted educators	0.152 (0.045)	3.407	< 0.001	0.065	0.240
Manualized fully	-0.119 (0.052)	-2.281	0.023	-0.221	-0.017
LO: Language	-0.392 (0.052)	-7.545	< 0.001	-0.494	-0.290
Targeted preschool	0.282 (0.078)	3.606	< 0.001	0.129	0.435
LO: Other	0.181 (0.041)	4.413	< 0.001	0.100	0.261
Age	-0.008 (0.005)	-1.692	0.091	-0.018	0.001
Increasing quality	-0.131 (0.043)	-3.056	0.002	-0.215	-0.047
Ingredient: Process quality	0.145 (0.049)	2.948	0.003	0.048	0.241

Interventions aimed at educators (teachers in preschool, school and daycare centres) are significantly more effective than those targeted directly at children or their parents (or others). Fully manualised interventions are less effective than those that are not manualised or only partially manualised. Language as a learning outcome (as opposed to maths) is associated with significantly lower effect sizes, while "other" outcomes (often vocabulary) have larger effects. Interventions targeted at kids below school age (preschool) are more effective than those targeting older children. With regard to the content of interventions, those intended to increase quality tend to have lower effects than those aimed at increasing quantity. However, interventions that focus on process quality rather than structural quality or learning formats are more effective.

Interestingly, the evaluation design (RCT vs other methods) does not matter for the effect size. This implies that, at least on average, non-experimental evaluation designs do not significantly over- or underestimate the true effect size (as measured by RCTs) in our sample of effects. However, this does not imply that individual non-RCT studies are free from bias, as there may be biases in different directions that cancel out on average.

Table 5 presents the results of the forward selection meta-regression for the cost-standardized effect sizes. Again, language as an outcome is significantly less cost-effective compared to maths or "other". Interventions targeting preschoolers are more cost-effective than those targeting older children and young adults, especially if the intervention takes place in a daycare setting. Interventions targeting educators are also more cost-effective, whereas fully manualised interventions are less cost-effective. Interestingly, on average, it does not matter whether the cost-standardised effect size was estimated using an RCT or another evaluation design. As such, many of the results from the meta-regressions on raw effect sizes carry over to cost-standardised effect sizes.

Table 5. Multivariate Meta-Regression Results – cost-standardised effect sizes

Variance Components					
Component	Estimate	Sqrt	N levels	Fixed	Factor
σ_4^2	0.000	0.002	31	No	Study.id
σ_3^2	0.000	0.002	48	No	Study.id/ treat.id
σ_2^2	0.000	0.001	61	No	Study.id/ treat.id/es.id
Test for Heterogeneity					
Test	Df	Q-value	p-value		
Cochran's Q	55	235.006	< 0.001		
Model Results					
	Estimate (SE)	z-value	p-value	CI (lower)	CI (upper)
Intercept	0.002 (0.001)	2.214	0.027	0.000	0.005
LO: Language	-0.169 (0.022)	-7.512	< 0.001	-0.213	-0.125
Targeted preschool	0.168 (0.024)	7.081	< 0.001	0.122	0.215
Takes place at daycare	0.073 (0.019)	3.828	< 0.001	0.036	0.110
Targeted educators	0.043 (0.008)	5.409	< 0.001	0.028	0.059
Manualized fully	-0.039 (0.009)	-4.519	< 0.001	-0.056	-0.022

4.2.1 Test for publication bias

Generally speaking, the prevalence of publication bias is a major concern in meta-analysis. If the probability of not identifying studies increases as their level of significance decreases, the result will be biased. We argue that the following three mechanisms diminish the likelihood of publication bias:

1. We include not only studies published in journals, but also studies published in public reports aimed at making policy recommendations.
2. Preregistration: To secure funding for interventions, TFCCR typically adheres to a preregistration process, which encourages publication regardless of whether the results are statistically significant.

3. TrygFonden's Centre for Child Research typically conducts an a priori power analysis to ensure that each intervention has a sufficient sample size to detect a statistically significant effect, based on the expected effect. Consequently, we expect relatively fewer non-rejections of a false null in our study population. This means that the proportion of studies that remain unpublished solely due to non-significant results should be smaller.

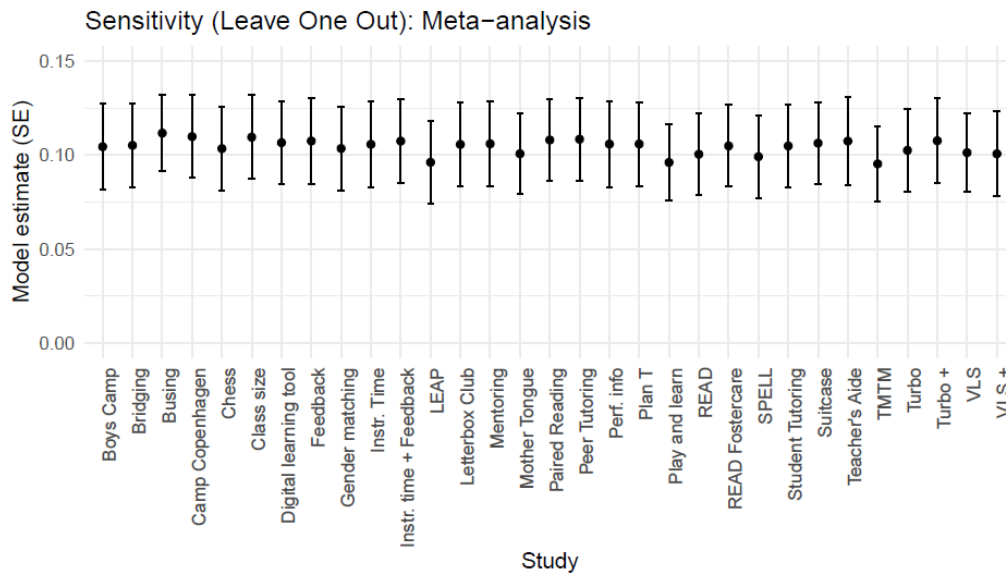
However, even though we argue that these mechanisms reduce publication bias in our study population, we still acknowledge that some bias may remain. Figure A.1 in the Appendix depicts the funnel plot for both raw and cost-standardised effect sizes. For the raw effect size, we see a relatively symmetric shape with higher precision, which is closer to the true effect size, and a symmetric distribution when standard errors become larger. If publication bias were present, we would expect small studies (with low precision) to be clustered on the right-hand side of the true effect size. For the effect size per \$100, we actually see some right-hand side clustering in a few studies. However, studies with a relatively large effect size compared to the pooled effect size are downscaled to a much smaller extent as they are relatively cheaper.

We use rank correlation to formally test for publication bias (following the approach of Begg and Mazumdar). This method consists of testing the positive correlation between effect sizes and standard errors, the results of which are shown in Table 6.

Table 6. Test for publication bias

	Kendall's τ	p-value
Raw effect size	0.058	0.5
Cost-standardised effect size	0.251	0.004

Figure 5. LOO sensitivity analysis for raw effect sizes – no predictors



When it comes to the raw effect size, it is evident that we cannot reject the null hypothesis of no publication bias. On the contrary, we reject the null hypothesis for the effect size per \$100 – but again, we argue that this is essentially a constructed estimate based on costs, which should not affect the probability of publishing.

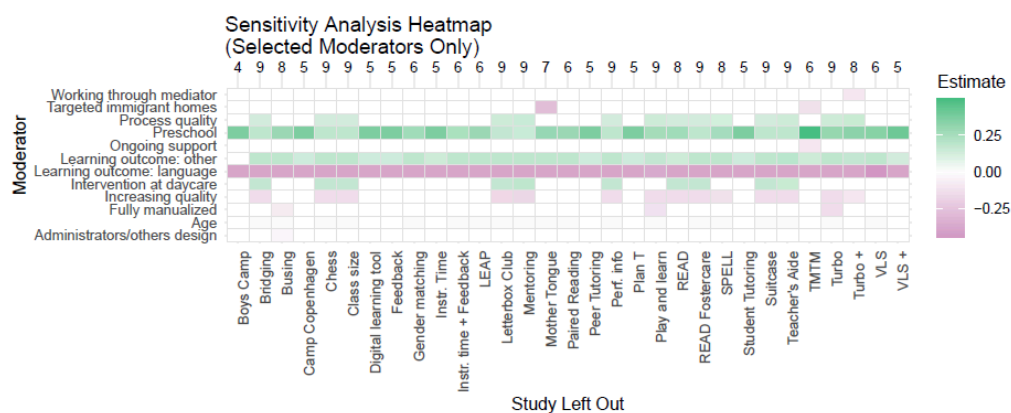
4.2.2 Investigation of influential studies

To investigate whether the effects are driven by a few highly influential studies, we perform a Leave-One-Out (LOO) sensitivity analysis.^[24] We exclude one study at a time and apply both the meta-analysis without predictors and the meta-regressions with predictors to see if the estimates change dramatically or are sensitive to the exclusion of a single study. Figure 5 visualises the results of the LOO analysis for the meta-analysis without predictors for raw effect sizes. The figure shows that the estimated overall effect size is both stable (approximately 0.1) and consistently statistically significant. The visualisation does not indicate the presence of any influential studies.

Figure 6 visualises the result of the LOO analysis for the meta-regressions with predictors. While the number of selected predictors can vary considerably between regression fits, a clear picture emerges of predictors that are consistently chosen with a similar size of estimate. We therefore find no indication of the presence of influential studies in the meta-regressions.

24. For information about the specific studies, see online SI.

Figure 6. LOO sensitivity analysis for raw effect sizes – meta-regression with predictors. Numbers above the graph indicate the number of selected predictors

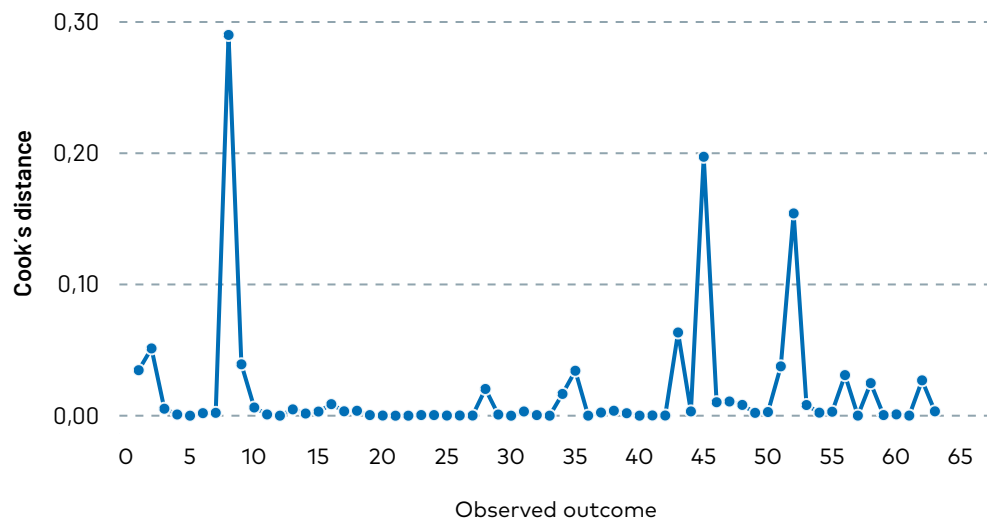


Appendix D presents the results of a sensitivity analysis of cost-standardised effect sizes. These indicate no cause for general concern about the robustness of the cost-standardised effect sizes.

We also use Cook's distance to explore potential outliers, as visualised in Figure 7. While these graphs indicate a few potentially influential results, the sensitivity analyses above suggest that the influences are negligible for the final estimates.

Figure 7. Cook's distances

Panel A: Raw effect sizes



Panel B: Cost-standardised effect sizes

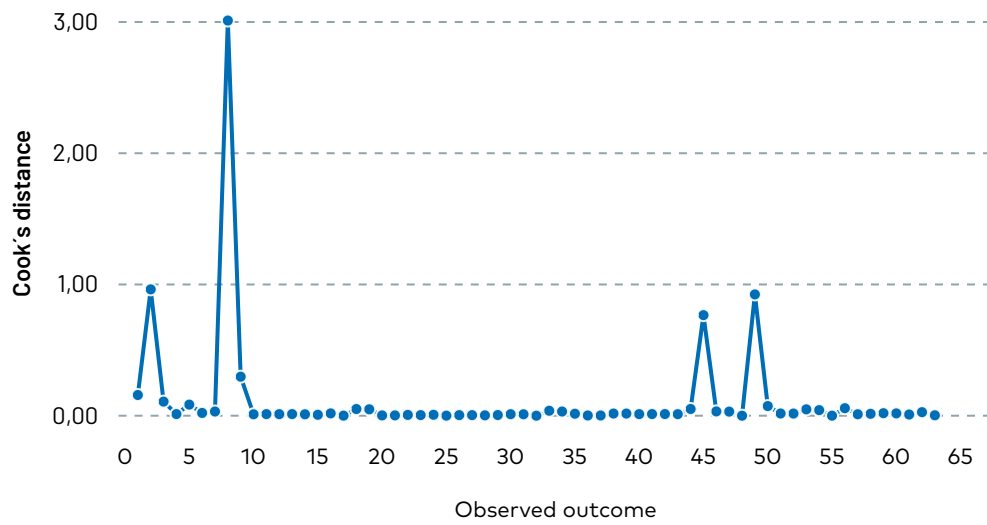
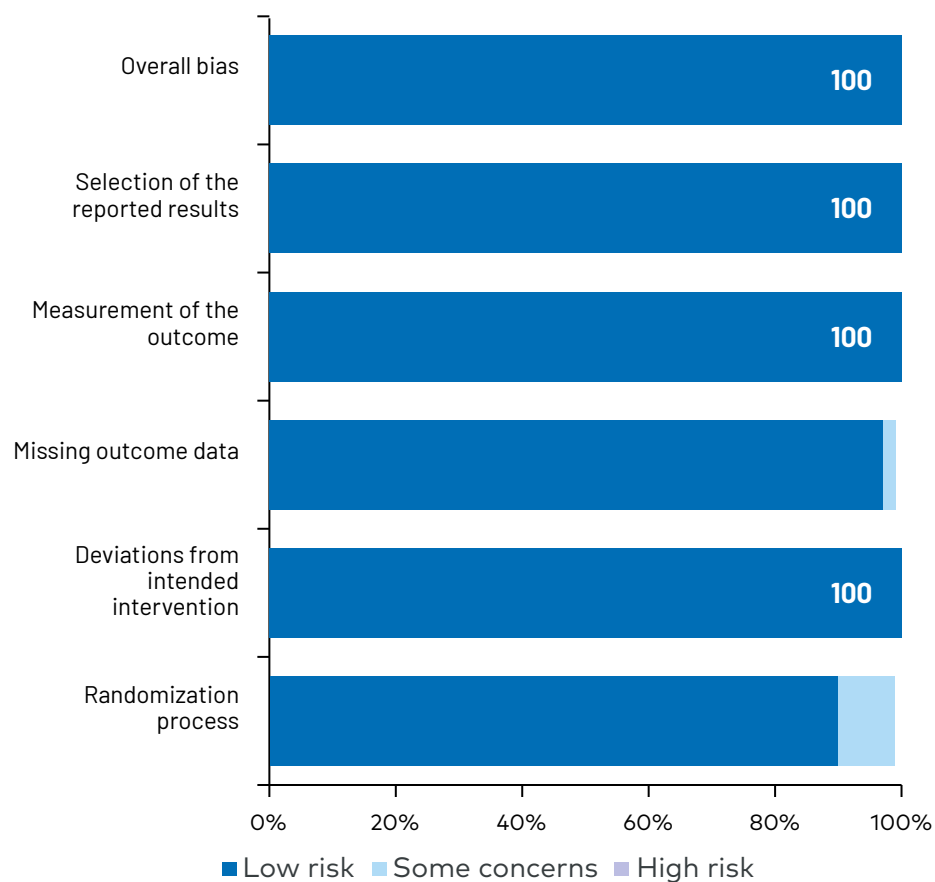


Figure 8. Risk of bias assessment for RCT studies

As percentage (intention-to-treat)



4.2.3 Risk-of-bias assessment

The risk-of-bias assessment determined that all of the studies that meet the inclusion criteria have an overall low risk of bias (see Figure 8 for visualisation of RCT studies).

The algorithm in the assessment tool concluded that several studies were of *some concern* in terms of risk of bias, due to a lack of preregistered analysis. However, as we are in a unique position to gather information about the study details from the authors concerned, we can safely conclude that the lack of preregistration (the most common cause of concern, especially in non-experimental studies) did not introduce particularly high risks of bias, since the analyses were carried out as planned and internally registered. Several studies present null findings, which supports the judgement that the studies did not change the analysis plan, nor did they select results to present, which would increase the risk of bias. Many studies use data from the national tests

available at Statistics Denmark as outcome measures. Since these tests are mandatory for students in public-sector schools in Denmark and are taken electronically, this strengthens the assessment's conclusion that there is a low risk of bias. Tables A.3 and A.4 in the appendix present an overview of risk of bias assessments for all studies (treatment arm/outcome measure), while the full assessments are available in the [online SI](#).

5 Discussion

The meta-analysis of the raw effect sizes found a reliable positive population effect size estimate of 0.104 (95% CI: 0.065 - 0.144), with 79% of the estimates being positive and 52% of those being significant. The meta-analysis of the cost-standardised effect sizes also found a positive population effect size estimate (ES: 0.037, 95% CI: 0.018 - 0.057), with 80% of the estimates being positive and 53% of those being significant.

The fact that most interventions have a positive effect indicates the success of TrygFonden's Centre for Child Research's thorough, scientific approach to creating and identifying effective interventions. While the interventions may differ in terms of their specific characteristics, they share a certain methodological rigor. The meta-analyses suggest that carefully designed interventions have a generally positive impact on various learning outcomes for children at all ages.

When unravelling the overall pattern, the interventions vary to a large extent in terms of their content and effect size. The most effective intervention measured in raw effect size is Play and Learn. This intervention is a 20-week, low-cost curriculum that runs in Danish childcare centres and trains teachers to apply sequenced weekly language and maths targets, engage in rich "serve and return" interactions and incorporate self-designed activities into everyday routines. The content is supported by planners, posters, logs and a two-day workshop. The second-most effective intervention is a treatment arm within the TMTM study consisting of a 12-week programme in which a trained mathematics teacher provides structured, dialogue-based one-to-one instruction four times per week, supported by screening, ongoing diagnostic conversations and targeted tasks. The teaching is continually adapted to each pupil's knowledge gaps, and the teacher receives three video-based supervision sessions from a school maths coach. This illustrates the significant differences between the two most effective interventions: i.e., aimed at language vs mathematics skills; and taking place in childcare vs school. All of the included interventions are described in more detail in the [online SI](#).

We conducted meta-regressions to investigate whether key characteristics are systematically associated with higher or lower effect sizes. The regressions on the raw effect sizes show that interventions targeting educators and interventions targeting

children at the preschool age are significantly more effective. This is consistent with similar findings in the literature that interventions targeted at younger children are more effective (see, e.g., Heckman (2006) and Rosholm et al. (2021)). Interventions with language as a targeted outcome, as well as fully manualised interventions, are found to be significantly less effective. While this latter finding is interesting, it should be interpreted in light of the fact that only 12.7% of studies include a fully manualised intervention. Lastly, while interventions that increase quality seem to be less effective than interventions that increase quantity, the meta-regression also indicates that a focus on process quality is significantly more effective than a focus on other aspects of quality.

When effect sizes are standardised by cost, the relative ranking changes. The most effective intervention then becomes Performance Information, which is evaluated for both maths and reading.

This is a quasi-experimental study that takes advantage of an unplanned IT breakdown during national tests to evaluate how students are affected by receiving information on their current performance. The theoretical basis for this is that knowledge of your own performance in itself has an impact on future performance. The second most effective cost-standardised intervention is the Feedback intervention. This is another quasi-experimental design, which takes advantage of the fact that the raw scores from national tests can be mapped mechanically onto five categories – from “considerably below average” to “considerably above average”. This means that students whose scores fall just below a given cutoff are seen as performing significantly worse than otherwise identical students whose scores are just above it. The study then investigates how this relatively negative feedback affects later achievements.

Common to these highly cost-effective interventions is that they were conducted in relation to national tests. In both cases, the active treatment is related to receiving feedback on performance, but the test is taken under all circumstances, which makes these interventions very cheap and therefore highly cost-effective. All of these interventions are described in the [online SI](#).

Formally, the meta-regression of the cost-standardised effect sizes identified a pattern in terms of the effect of different mediators and moderators that was very similar to the pattern for raw effect sizes. Again, early interventions (which target children in preschool and take place at daycare) and interventions targeting educators are more effective, while fully manualised interventions and interventions in which language is the learning outcome are less effective.

We found no clear correlation between raw effect sizes and intervention costs. This is surprising, as we expected more expensive interventions to have a larger effect. Our

assumption was that more expensive interventions would be associated with either higher quality or quantity, and in both cases, we would expect an increase in the treatment effect. The fact that this pattern was *not* found in the analysis may be partly explained by the fact that the most effective cost-standardised interventions were quasi-experimental studies, in which the treatment consisted of feedback on national test performance. However, the leave-one-out analysis indicated that no individual study was driving the effects. Further, significant findings in the meta-regressions identified other characteristics associated with higher effect sizes. For example, we found that interventions focusing on increasing quality were less effective. However, this does not necessarily imply that higher quality is not effective – it may also reflect more general difficulties associated with influencing and enhancing quality rather than quantity. This result also stands in contrast to the finding that increasing process quality as an active ingredient is associated with larger effects.

Interestingly, neither of the two meta-regressions found that an RCT design is a significant predictor of effect size. This finding indicates that non-RCT approaches are, on average, unbiased – at least in our sample. This is interesting, since it suggests that combining RCT studies with other carefully designed evaluations may provide insight into treatment effects in various settings. This finding may also reflect the high-quality of Danish register data, which allows researchers to take advantage of naturally occurring exogenous variation. A recent study by Imbens & Xu (2025) directly compares RCTs with quasi-experimental studies. It finds that while non-RCTs are sensitive to specification and only identify causal relationships when relatively strong assumptions hold, they can also, given the right circumstances, present an alternative to experimental designs. This specification sensitivity, along with the relatively low sample size of the present analysis, means we cannot conclude anything about the different methodological approaches. We can investigate treatment effects in types of interventions, for which only a small amount of data has been collected, as long as we make an effort to meet assumptions. However, due to restrictions and highly data-dependent assumptions, quasi-experimental methods cannot be used exclusively for evaluations of education-type interventions – and in cases where RCTs are feasible, these are always preferable.

The complex pattern of intervention characteristics^[25] may also serve as a starting point for further intervention development. In general, carefully designed interventions have a positive effect on various learning outcomes, but their specific effect sizes vary. As the costs of interventions were found not to be important for treatment effects, it may be more fruitful to focus on content and structural parameters when designing new interventions.

25. For details of all studies, see the online SI.

6 Limitations

There are several limitations that influence the extent to which these conclusions can be generalised. As in Rosholm et al. (2021), the meta-analysis only includes short-term effects, which may not necessarily persist in the long term. The effect of an intervention may fade out over time (and the degree of fade-out can vary between studies), but it could also be reinforced if future investments become more efficient as a result.

The costs of the treatments are treated as fixed. In practice, however, some errors invariably arise when measuring intervention costs. This should be taken into consideration when interpreting the findings on cost-standardised effects. In order to minimise errors, all cost-standardised effects are reported in 2021 prices and, as such, take inflation into account.

We consider the internal validity of this study to be high, since all interventions were conducted in the same country, around the same time period (2013–2024). However, it is important to note that many interventions were implemented at the classroom (or group) level, and as such, capture not only the direct impact on individual children, but also potential spillover effects from other children within the treated classroom. It is possible that spillover effects within classrooms serve as a significant mechanism through which interventions operate. While this should be considered when interpreting the absolute size of treatment effects, it is less relevant for comparisons across studies, as most studies are implemented at the classroom level.

Regarding external validity, we believe the results are most applicable to other Scandinavian welfare states. In other countries with a lower level of spending on children, particularly in daycare settings, interventions of the type included here may in fact yield even larger effect sizes.

As stated, this study only includes studies from one research centre. Studies from all of the Nordic countries provide valuable insights into educational research and policy-making. However, the limitations of this meta-analysis – including a lack of cost information and data on mediators and moderators – make it impossible to include all Nordic intervention studies. This approach allows for a richer set of insights, but it also excludes potentially insightful studies, and as such may not accurately reflect the overall effectiveness of interventions in the region. By only including studies from one research centre, we may miss out on potential insights and a broader understanding of the overall effectiveness of interventions in Denmark and the Nordic countries. This may lead to a skewed perspective on the overall effectiveness of interventions, which may not accurately reflect all interventions in the Nordic Region.

7 Conclusion

This paper builds upon and extends the meta-analysis by Rosholm et al. (2021) by including more RCTs and non-RCTs conducted in Denmark in collaboration with TrygFonden's Centre for Child Research. We use a four-level multivariate meta-model to investigate intervention treatment effects and find a positive overall estimated population effect. This is the case for both raw effect sizes and cost-standardised effect sizes. We use meta-regressions to investigate the impact of the interventions' key characteristics, and this approach replicates findings in the literature that interventions are more effective when targeted at younger children, which supports the theoretical implications of the Heckman-curve theory.

Surprisingly, we see no correlation between the intervention cost and effect size, which may be caused by general difficulties associated with increasing quality in education-type interventions. We also find that evaluation design (RCT vs non-RCT) is not a significant predictor of effect size. This indicates that, in our sample of studies, non-RCT evaluation methods are unbiased on average and may function well as alternatives to RCT studies where the latter are not feasible.

Lastly, we discuss the limitations of our study, including measurement errors on costs of interventions, internal and external validity and regional limitations.

References

- Andersen, S. C., Beuchert, L., & Nielsen, H. S. (2024, May). The effect of teacher's aides on students with special needs. <https://doi.org/10.2139/ssrn.4182253>
- Andersen, S. C., Beuchert, L., Nielsen, H. S., & Thomsen, M. K. (2020). The effect of teacher's aides in the classroom: Evidence from a randomized trial. *Journal of the European Economic Association*, 18(1), 469–505. <https://doi.org/10.1093/jeaa/jvy048>
- Andersen, S. C., Guul, T. S., & Humlum, M. K. (2021). How first-language instruction transfers to majority-language skills. *Nature Human Behaviour*, 6(2), 229–235. <https://doi.org/10.1038/s41562-021-01200-x>
- Balduzzi, S., Rücker, G., & Schwarzer, G. (2019). How to perform a meta-analysis with R: A practical tutorial. *Evidence-Based Mental Health*, 22(4), 153–160. <https://doi.org/10.1136/ebmental-2019-300117>
- Barro, R. J. (2013). Education and economic growth. *Annals of Economics and Finance*, 14(2), 301–328.
- Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, 50(4), 1088–1101. <https://doi.org/10.2307/2533446>
- Bleses, D., Horn, N. T., Larsen, E. N. F., Larsen, N. S., Munkedal, S., Thingholm, P. R., & Wandall, J. (2022, September). *Evaluering af obligatoriske sprogprøver* (Technical report). Trygfondens Børneforskningscenter, Epinion, Aarhus University.
- Burchinal, M., Vandergrift, N., Pianta, R., & Mashburn, A. (2010). Threshold analysis of association between child care quality and child outcomes for low-income children in pre-kindergarten programs. *Early Childhood Research Quarterly*, 25(2), 166–176. <https://doi.org/10.1016/j.ecresq.2009.10.004>
- Byrne, B., & Fielding-Barnsley, R. (1991). Evaluation of a program to teach phonemic awareness to young children. *Journal of Educational Psychology*, 83(4), 451–457.
- Campbell, F. A., Ramey, C. T., Pungello, E., Sparling, J., & Miller-Johnson, S. (2002). Early childhood education: Young adult outcomes from the Abecedarian Project. *Applied Developmental Science*, 6(1), 42–57. <https://doi.org/10.1207/S1532480XADS060105>
- Chancel, L., & Piketty, T. (2021). Global income inequality 1820–2020: The persistence and mutation of extreme inequality. *Journal of the European Economic Association*, 19(6), 3025–3062. <https://doi.org/10.1093/jeaa/jvab047>
- Clements, D. H., & Sarama, J. (2007). Effects of a preschool mathematics curriculum: Summative research on the Building Blocks project. *Journal for Research in Mathematics Education*. <https://doi.org/10.2307/30034954>

- Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics*, 10, 417–451.
- Duncan, G. J., & Magnuson, K. (2013). Investing in preschool programs. *Journal of Economic Perspectives*, 27(2), 109–132. <https://doi.org/10.1257/jep.27.2.109>
- Ehri, L. C., Nunes, S. R., Stahl, S. A., & Willows, D. M. (2001). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research*, 71(3), 393–447. <https://doi.org/10.3102/00346543071003393>
- Elango, S., García, J. L., Heckman, J. J., & Hojman, A. (2016). Early childhood education. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.), *Economics of means-tested transfer programs in the United States, volume II* (pp. 235–298). University of Chicago Press. <https://doi.org/10.7208/9780226392523-005>
- Fredriksson, P., Öckert, B., & Oosterbeek, H. (2013). Long-term effects of class size. *Quarterly Journal of Economics*, 128(1), 249–285. <https://doi.org/10.1093/qje/qjs048>
- Fryer, R. G. (2017). The production of human capital in developed countries: Evidence from 196 randomized field experiments. In A. V. Banerjee & E. Duflo (Eds.), *Handbook of economic field experiments* (Vol. 2, pp. 95–322). North-Holland. <https://doi.org/10.1016/bs.hefe.2016.08.006>
- Gardner, F., Leijten, P., Melendez-Torres, G., Landau, S., Harris, V., Mann, J., Beecham, J., Hutchings, J., & Scott, S. (2019). The earlier the better? Individual participant data and traditional meta-analysis of age effects of parenting interventions. *Child Development*, 90(1), 7–19. <https://doi.org/10.1111/cdev.13138>
- Haghighi, E., Vach, W., Højen, A., & Bleses, D. (2021). Estimating measurement error in child language assessments administered by daycare educators in large-scale intervention studies. *PLOS ONE*, 16(11). <https://doi.org/10.1371/journal.pone.0304884>
- Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2003). Does peer ability affect student achievement? *Journal of Applied Econometrics*, 18(5), 527–544. <https://doi.org/10.1002/jae.741>
- Hanushek, E. A., & Woessmann, L. (2010). Education and economic growth. *Economics of Education*, 60, 1–67.
- Hattie, J. (2023). *Visible learning: The sequel*. Routledge.
- Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, 312(5782), 1900–1902. <https://doi.org/10.1126/science.1128898>

- Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11), 1539–1558. <https://doi.org/10.1002/sim.1186>
- Hoff, E., Trecca, F., Højen, A., Laursen, B., & Bleses, D. (2024). Context and education affect the quality of parents' speech to children. *Journal of Applied Developmental Psychology*, 91, 101632. <https://doi.org/10.1016/j.appdev.2024.101632>
- Højen, A., Hoff, E., Bleses, D., & Dale, P. S. (2021). The relation of home literacy environments to language and preliteracy skills in single- and dual-language children in Danish childcare. *Early Childhood Research Quarterly*, 55, 312–325. <https://doi.org/10.1016/j.ecresq.2020.12.007>
- Højen, A., Madsen, T. O., & Bleses, D. (2023). Danish 20-month-olds' recognition of familiar words with and without consonant and vowel mispronunciations. *Phonetica*, 80(5), 309–328. <https://doi.org/10.1515/phon-2023-2001>
- Humlum, M. K., & Smith, N. (2015). Long-term effects of school size on students' outcomes. *Economics of Education Review*, 45, 28–43. <https://doi.org/10.1016/j.econedurev.2015.01.003>
- Hvidman, C., Koch, A., Nafziger, J., Nielsen, S. A., & Rosholm, M. (2020). An intensive, school-based learning camp targeting academic and non-cognitive skills evaluated in a randomized trial. *IZA Discussion Paper*.
- Imbens, G. W., & Rubin, D. B. (2015). *Causal inference for statistics, social, and biomedical sciences*. Cambridge University Press.
- Imbens, G. W., & Xu, Y. (2025). Comparing experimental and non-experimental methods: What lessons have we learned four decades after Lalonde (1986)? *Journal of Economic Perspectives*, 39(4), 173–201.
- Jensen, P., & Sjö, N. M. (2024). The effects of a large-scale school readiness intervention on Danish preschool children's emergent mathematics skills. *Scandinavian Journal of Educational Research*, 68(3), 488–503. <https://doi.org/10.1080/00313831.2022.2148281>
- Jepsen, C., & Rivkin, S. (2009). Class size reduction and student achievement. *Journal of Human Resources*, 44(1), 223–250. <https://doi.org/10.3368/jhr.44.1.223>
- Justice, L. M., McGinty, A. S., Piasta, S. B., Kaderavek, J. N., & Fan, X. (2010). Print-focused read-alouds in preschool classrooms. *Language, Speech, and Hearing Services in Schools*, 41(4), 504–520. [https://doi.org/10.1044/0161-1461\(2010/09-0056\)](https://doi.org/10.1044/0161-1461(2010/09-0056))
- Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Bill & Melinda Gates Foundation.

Levitt, C. (2009). *From best practices to breakthrough impacts: A science-based approach to building a more promising future for young children and families*. Center on the Developing Child, Harvard University.

Lundahl, B. W., Tollefson, D., Risser, H., & Lovejoy, M. C. (2008). A meta-analysis of father involvement in parent training. *Research on Social Work Practice, 18*(2), 97–106. <https://doi.org/10.1177/1049731507309828>

Nielsen, S. A. (2023). Boys left behind: The effects of summer camp and follow-up strategies on academic, personal, and social competencies. *Economics of Education Review, 93*, 102370. <https://doi.org/10.1016/j.econedurev.2023.102370>

OECD. (2024). *Society at a glance 2024: OECD social indicators*. <https://doi.org/10.1787/918d8db3-en>

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. National Institute of Child Health and Human Development.

Phillips, D. A., Lipsey, M. W., Dodge, K. A., Haskins, R., Bassok, D., Burchinal, M. R., & Weiland, C. (2017). Puzzling it out: The current state of scientific knowledge on pre-kindergarten effects. In *The current state of scientific knowledge on pre-kindergarten effects* (pp. 19–30).

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.r-project.org/>

Reynolds, A. J., Temple, J. A., Ou, S.-R., Robertson, D. L., Mersky, J. P., Topitzes, J. W., & Niles, M. D. (2007). Effects of a school-based early childhood intervention on adult health and well-being: A 19-year follow-up of low-income families. *Archives of Pediatrics & Adolescent Medicine, 161*(8), 730–739. <https://doi.org/10.1001/archpedi.161.8.730>

Rosholm, M., Paul, A., Bleses, D., Højen, A., Dale, P. S., Jensen, P., Justice, L. M., Svarer, M., & Andersen, S. C. (2021). Are impacts of early interventions in the Scandinavian welfare state consistent with a Heckman curve? A meta-analysis. *Journal of Economic Surveys, 35*(1), 106–140. <https://doi.org/10.1111/joes.12400>

Rosholm, M., Tønnesen, P. B., Rasmussen, K., Overgaard, S., Færch, J. V., Malm, S. G., & Harder, J. (2025). A tailored small group instruction intervention in mathematics benefits low achievers. *npj Science of Learning, 10*(1), 18. <https://doi.org/10.1038/s41539-025-00310-9>

- Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomized trials. *BMJ*, 340, c332. <https://doi.org/10.1136/bmj.c332>
- Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H.-Y., Corbett, M. S., Eldridge, S. M., et al. (2019). RoB 2: A revised tool for assessing risk of bias in randomized trials. *BMJ*, 366.
- Svane, R. P., Willemsen, M. M., Bleses, D., Krøjgaard, P., Verner, M., & Nielsen, H. S. (2023). A systematic literature review of math interventions across educational settings from early childhood education to high school. *Frontiers in Education*, 8. <https://doi.org/10.3389/feduc.2023.1229849>
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3), 1–48. <https://doi.org/10.18637/jss.v036.i03>
- White, L. A., Prentice, S., & Perlman, M. (2015). The evidence base for early childhood education and care programme investment: What we know, what we don't know. *Evidence & Policy*. <https://doi.org/10.1332/174426415X14210818992588>

Appendix

A Table with data summary

Table A.2: Data Summary

Study Specification		Attributes				
Study	Treatment	Domain	Effect Size	Standard Error	Cost	RCT
Play and learn (Bleses, Jensen, et al, 2020)	VLS 1-2 Years –Centre Based	Math	0.68	0.12	232	X
Play and learn (Bleses, Jensen, et al, 2020)	VLS 1-2 Years – Centre Based	Language	0.23	0.07	232	X
SPELL (Bleses, Højen, et al., 2018b; Bleses et al., 2015)	SPELL (Baseline)	Language	0.27	0.08	92	X
SPELL (Bleses, Højen, et al., 2018b; Bleses et al., 2015)	SPELL + PD	Language	0.28	0.14	130	X
SPELL (Bleses, Højen, et al., 2018b; Bleses et al., 2015)	SPELL + HOME	Language	0.18	0.05	110	X
LEAP (Bleses, Højen, et al., 2018a; Bleses et al., 2015)	LEAP LARGE	Language	0.04	0.06	69	X
LEAP (Bleses, Højen, et al., 2018a; Bleses et al., 2015)	LEAP SMALL	Language	0.06	0.07	96	X
LEAP (Bleses, Højen, et al., 2018a; Bleses et al., 2015)	LEAP OPEN	Language	0.43	0.07	101	X
READ (S. C. Andersen & Nielsen, 2016)	READ	Read	0.26	0.07	83	X
Suitcase (Jakobsen & Andersen, 2013)	Suitcase	Language	-0.08	0.21	234	X
Teacher's Aide (S. Andersen et al., 2018)	With Degree	Read	0.09	0.03	1181	X
Teacher's Aide (S. Andersen et al., 2018)	Without Degree	Read	0.13	0.04	1251	X
Teacher's Aide (S. Andersen et al., 2018)	With Degree	Math	0.06	0.05	1181	X
Teacher's Aide (S. Andersen et al., 2018)	Without Degree	Math	0.04	0.05	1251	X
Instr. Time (S. C. Andersen et al., 2016)	Without Teaching Programme	Read	0.15	0.06	201	X
Instr. Time (S. C. Andersen et al., 2016)	With Teaching Programme	Read	0.04	0.06	201	X
Chess (Rosholm et al., 2017)	Chess Instructions	Math	0.16	0.08	301	
Mentoring (Svarer et al., 2014)	Mentors	Education	0.07	0.04	3900	X

Bridging (Rosholm, Mikkelsen, & Svarer, 2019)	Bridging	Education	0.10	0.01	2537	
TMTM (Rosholm et al., 2025)	Teacher to Group – Grade 2	Math	0.48	0.20	396	X
TMTM (Rosholm et al., 2025)	Teacher to Group – With Supervision – Grade 2	Math	0.09	0.21	552	X
TMTM (Rosholm et al., 2025)	Teacher to Student – With Supervision – Grade 2	Math	0.63	0.22	1284	X
TMTM (Rosholm et al., 2025)	Teacher to Group of High Achieving Students – Grade 2	Math	0.21	0.12	460	X
TMTM (Rosholm et al., 2025)	Teacher to Group of High Achieving Students – With Supervision – Grade 2	Math	0.28	0.13	674	X
TMTM (Rosholm et al., 2025)	Teacher to Group – Grade 8	Math	0.29	0.25	483	X
TMTM (Rosholm et al., 2025)	Teacher to Group – With Supervision- Grade 8	Math	0.49	0.24	616	X
TMTM (Rosholm et al., 2025)	Teacher to Student – With Supervision – Grade 8	Math	0.34	0.26	919	X
Turbo (Rosholm, Hvidman, et al., 2020)	Turbo	Read	0.01	0.22	795	X
Turbo (Rosholm, Hvidman, et al., Turbo 2020)	Turbo	Math	0.19	0.08	795	X
Turbo + (Hvidman et al., 2020)	Turbo+	Read	0.00	0.03	786	X
Turbo + (Hvidman et al., 2020)	Turbo+	Math	0.06	0.04	786	X
Perf. info (S. Andersen & Nielsen, 2020)	Performance information	Read	0.09	0.02	7	
Perf. info (S. Andersen & Nielsen, 2020)	Performance information	Math	0.07	0.05	7	
Instr. time + Feedback (S. Andersen et al., 2019)	Instr. Time and Feedback	Read	0.02	0.04	199	X
Mother Tongue (S. C. Andersen et al., 2022)	Mother Tongue Instruction	Read	0.46	0.18	1419	X
Feedback (Beuchert et al., 2020)	Feedback On Tests	Read	0.03	0.01	7	
Feedback (Beuchert et al., 2020)	Feedback On Tests	Math	0.06	0.02	7	
Camp Copenhagen (Rosholm, Nielsen, Hvidman, et al., 2020)	Camp	Read	-0.02	0.05	6234	
Camp Copenhagen (Rosholm, Nielsen, Hvidman, et al., 2020)	Camp	Math	-0.04	0.06	6234	
Boys Camp (Rosholm, Nielsen, Hvidman, et al., 2020)	Camp	Math	-0.04	0.04	5994	

Boys Camp (Rosholm, Nielsen, Hvidman, et al., 2020)	Camp	Read	-0.04	0.04	5994	
Plan T (Rosholm, Nielsen, Hvidman, et al., 2020)	Camp	Read	0.07	0.06	4824	
VLS + (Bleses, Moos, & Boisen, 2020) (Bleses et al., 2025)	VLS 1-2 Years – Centre Based	Language	0.50	0.12	228	X
VLS + (Bleses, Moos, & Boisen, 2020) (Bleses et al., 2025)	VLS And Leaders	Language	0.02	0.04	325	X
VLS + (Bleses, Moos, & Boisen, 2020) (Bleses et al., 2025)	VLS And Leaders And Staff	Language	0.00	0.04	379	X
Digital learning tool (Gissel & Andersen, 2021)	Digital learning tool	Read	0.03	0.09	7	X
Class size (Nandrup, 2016)	Class size	Read	0.00	0.01	0	
Class size (Nandrup, 2016)	Class size	Math	0.00	0.01	0	
VLS (Bleses et al., 2021) Based	VLS 1-2 Years – Centre And Family Based	Language	0.09	0.04	231	X
VLS (Bleses et al., 2021)	VLS 1-2 Years – Centre and Family Based	Math	0.46	0.07	231	X
Busing (Damm et al., 2021)	Bussing	Read	-0.17	0.10	8815	
Busing (Damm et al., 2021)	Bussing	Math	-0.25	0.10	8815	
Gender matching (Guul, 2018)	Gender matching	Other	0.16	0.09	3900	
Peer Tutoring (Rosholm, Vind, et al., 2019)	Peer Tutoring	Read	-0.02	0.05	175	X
Student Tutoring (Rosholm, Vind, et al., 2019)	Student Tutoring	Math	0.08	0.27	210	X
Paired Reading (Rosholm, Vind, et al., 2019)	Paired Reading	Read	-0.14	0.15	1215	X
Letterbox Club (Rosholm, Vind, et al., 2019)	Club Pencil Case	Math	0.02	0.14	895	X
Letterbox Club (Rosholm, Vind, et al., 2019)	Club Pencil Case	Read	0.10	0.09	895	X
READ Foster Care (Bleses et al., 2022)	READ in Foster Care	Language	0.05	0.39	866	X
Boys Camp (Nielsen, 2023)	Boys Camp – Individual Follow Up	Other	0.12	0.05	5994	
Boys Camp (Nielsen, 2023)	Boys Camp – Group Follow Up	Other	0.20	0.04	5994	
Play and Learn in Kindergarten (Bleses et al., 2024)	Play and Learn in Kindergarten	Language	-0.02	0.07	102	X
Play and Learn in Daycare (Bleses et al., 2024)	Play and Learn in Daycare	Language	0.07	0.07	203	X

B Mathematical notation

B.1 4-level random-effects model

Following the notation in Harrer et al. (2021), the 4-level random-effects model is given by:

$$\hat{\theta}_{itk} = \mu + \zeta_{(4)k} + \zeta_{(3)tk} + \zeta_{(2)itk}$$

Where $\hat{\theta}_{itk}$ is the estimate of the true effect size i nested in treatment arm t within the study k . The random effect components $\zeta_{(2)itk}$ and $\zeta_{(3)tk}$ reflect the within-cluster heterogeneity on levels 2 and 3, respectively. $\zeta_{(4)k}$ reflects the between-cluster heterogeneity on level 4. μ is the parameter of interest, as it reflects the true population effect size. For $\zeta_{(2)itk} = \zeta_{(3)tk} = \zeta_{(4)k} = 0$, the model collapses to a fixed effect model, where all included effect sizes are drawn from the true population distribution of effect sizes. The sampling error ϵ_{itk} reflects the sampling error of the effect size. This quantity is assumed to be independent of the random components.

B.2 Formal specifications of heterogeneity and variance

The total heterogeneity estimate $\hat{\tau}^2$ is given by the sum of the variances from the random components: $\hat{\tau}^2 = \hat{\sigma}_4^2 + \hat{\sigma}_3^2 + \hat{\sigma}_2^2$

The proportion of total variance in the nested structure accounted for by heterogeneity is given by: $I^2 = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \hat{\sigma}_\epsilon^2}$

B.3 4-level mixed-effects model

The mixed-effects model applied at the meta-regressions is given by:

$$\hat{\theta}_{itk} = \mu + yx_{itk} + \zeta_{(4)k} + \zeta_{(3)tk} + \zeta_{(2)itk} + \epsilon_{itk}$$

where x_{itk} denotes the included mediators and moderators.

C Funnel plots for publication assessment

D Sensitivity analysis for cost-standardised effect sizes

Figures A.2 and A.3 below show the sensitivity analysis for the cost-standardised effect sizes. The moderator *Learning outcome: other* has an usually high estimate for the

analysis in which "Paired reading" is omitted. This behaviour seems to be an outlier, as all other results seem to have robust methods of leaving out specific studies.

Figure A.1: Funnel plot for both effect sizes

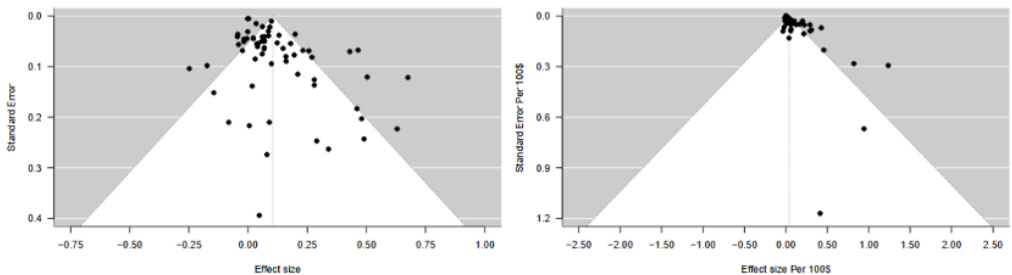


Figure A.2: LOO sensitivity analysis for cost-standardised effect sizes – no predictors

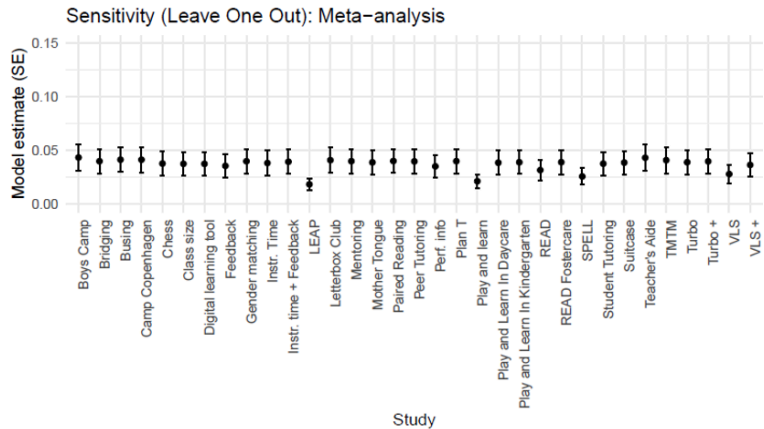
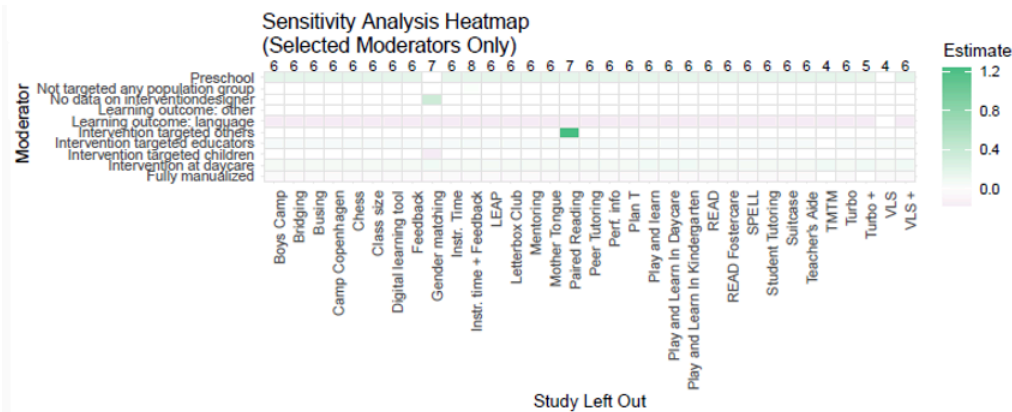


Figure A.3: LOO sensitivity analysis for cost-standardised effect sizes – meta-regression with predictors. Numbers above the graph indicate number of selected predictors



E Risk of bias assessment

Table A.3: Overview of risk of bias assessment for RCT studies. Both algorithm and assessor judgment is presented for each of the five domains.

The last judgment (bold) is the overall judgment of risk of bias. L = Low risk of bias, SC = some concerns for risk of bias, H = high risk of bias. Full assessments of each individual study and outcome are available in online SI.

Risk of Bias ID	Study	Outcome	Algorithm judgment	Assessor's judgment
bleses2020lowVocab	Play and learn	Vocabulary	L L L L SC L	L L L L L L
bleses2020lowMathskill	Play and learn	Math skills	L L L L SC L	L L L L L L
bleses2018SPELL	SPELL	Language assessment	L L SC L L L	L L L L L L
bleses2018SPELLHOME	SPELL + HOME	Language assessment	L L SC L L L	L L L L L L
bleses2018SPELLPD	SPELL + PD	Language assessment	L L SC L L L	L L L L L L
bleses2015LEAPSMALL	LEAP SMALL	Language overall	L L L L SC L	L L L L L L
bleses2015LEAPLARGE	LEAP SMALL	Language overall	L L L L SC L	L L L L L L
bleses2015LEAPOPEN	LEAP SMALL	Language overall	L L L L SC L	L L L L L L
andersen2016-Readingtest	Growth mindset	Reading test	L L L L SC SC	L L SC L L L
jakobsen2013-coproduction	Suitcase	Language overall	L L L L SC L	L L L L L L
andersen2018TA-WdegreeRead	TA with degree	Reading overall	L L L L SC L	L L L L L L
andersen2018TA-WdegreeMath	TA with degree	Math overall	L L L L SC L	L L L L L L
andersen2018TA-WOdegreeRead	TA without degree	Reading overall	L L L L SC L	L L L L L L
andersen2018TA-WOdegreeMath	TA without degree	Math overall	L L L L SC L	L L L L L L
andersen2016instr-Wprogram	Instruction time w program	Reading overall	L L L L SC L	L L L L L L
andersen2016instr-WOprogram	Instruction time w/o program	Reading overall	L L L L SC L	L L L L L L

Risk of Bias ID	Study	Outcome	Algorithm judgment	Assessor's judgment
svarer2014-Mentoring	Mentoring	Average education	L L L L SC SC	L SC L L L L
harder2020TMTM-TTG G2	TMTM (grade 2) teacher to group	Mathematics test	L L L L L L	L L L L L L
harder2020TMTM-TTG Wsuper G2	TMTM (grade 2) teacher to group w supervision	Mathematics test	L L L L L L	L L L L L L
harder2020TMTM-TTS Wsuper G2	TMTM (grade 2) teacher to student w supervision	Mathematics test	L L L L L L	L L L L L L
harder2020TMTM-TTG HighAch G2	TMTM (grade 2) teacher to group high achievers	Mathematics test	L L L L L L	L L L L L L
harder2020TMTM-TTG HighAch-W super G2	TMTM (grade 2) teacher to group high achievers w supervision	Mathematics test	L L L L L L	L L L L L L
harder2020TMTM-TTG G8	TMTM (grade 8) teacher to group	Mathematics test	L L L L L L	L L L L L L
harder2020TMTM-TTG Wsuper G8	TMTM (grade 8) teacher to group w supervision	Mathematics test	L L L L L L	L L L L L L
harder2020TMTM-TTS Wsuper G8	TMTM (grade 8) teacher to student w supervision	Mathematics test	L L L L L L	L L L L L L
rosholm2020-TurboReadingTest	Turbo	Reading overall	H L L L L SC	SC L L L L L
rosholm2020-TurboMathTest	Turbo	Reading overall	H L L L L SC	SC L L L L L
rosholm2020-Turbo+ReadingTest	Turbo+	Reading overall	H L L L L SC	SC L L L L L
andersen2022Mother-tongueReading	Mothertongue	Reading overall	L L L L SC L	L L L L L L
bleses2020VLS+-Language	VLS+	Language test	L L L SC SC L	L L L L L L

Risk of Bias ID	Study	Outcome	Algorithm judgment	Assessor's judgment
bleses2020VLS+- LeadersLanguage	VLS+ leaders	Language test	L L L SC SC L	L L L L L L
bleses2020VLS+- LeadersStaffLanguage	VLS+ leaders & staff	Language test	L L L SC SC L	L L L L L L
andersen2021- digitalRead	Digital lear- ning tool	Reading average	L L L L SC L	L L L L L L
bleses2021- VLSLanguage	VLS	Language average	L L L L SC L	L L L L L L
bleses2021- VLSLanguage	VLS	Math average	L L L L SC L	L L L L L L
rosholm2019- PeerReading	Peer tutoring	Reading overall	L L L L SC L	L L L L L L
rosholm2019- StudentReading	Student tutoring	Reading overall	L L L L SC L	L L L L L L
rosholm2019- ClubPencilReading	Club pencil case	Reading overall	L L L L SC L	L L L L L L
rosholm2019- ClubPencilMath	Club pencil case	Math overall	L L L L SC L	L L L L L L
bleses2022Read- FosterLanguage	READ in foster care	Language average	L L L L SC L	L L L L L L
bleses2022play- learnLanguage	Play & learn	Language assessment	L L L L SC L	L L L L L L
bleses2024play- learnLanguage	Play & learn	Language assessment	L L L L SC L	L L L L L L
andersen2019- instrTimeFeed-	Instruction time feedback	Reading overall	L L L L SC L	L L L L L L

Table A.4: Overview of risk of bias assessment for non-RCT studies. Both algorithm and assessor judgment is presented for each of the five domains. The last judgment (bold) is the overall judgment of risk of bias. L = Low risk of bias, SC = some concerns for risk of bias, H = high risk of bias. Full assessments of each individual study and outcome are available in online SI. File suffix specifies the file containing the full assessment. Filenames in the drive folder are on the form "ROBINS-I-V2 suffix"

File suffix Study	Outcome Algorithm judgment	Assessor's judgment
boysCamp Boys camp	Reading overall L L L L L L L L L	L L L L L L L L L
boysCamp Boys camp	Math overall L L L L L L L L L	L L L L L L L L L
boysCampV2 Boys camp (w individual follow-up)	GPA L L L L L L L L L	L L L L L L L L L
boysCampV2 Boys camp (w group follow-up)	GPA L L L L L L L L L	L L L L L L L L L
busing School busing	Reading overall L L L L L L L L L	L L L L L L L L L
busing School busing	Math overall L L L L L L L L L	L L L L L L L L L
camp- Copenhagen Camp Copenhagen	Reading overall L L L L L L L L L	L L L L L L L L L
camp- Copenhagen Camp Copenhagen	Math overall L L L L L L L L L	L L L L L L L L L
bridging Bridging	Completion of track L L L L L L L L L	L L L L L L L L L
chess Chess	Math test L L L L L L L L L	L L L L L L L L L
classSize Class size	Reading overall L L L L L L L L L	L L L L L L L L L
classSize Class size	Math overall L L L L L L L L L	L L L L L L L L L
feedback Feedback	Reading average L L L L L L L L L	L L L L L L L L L
feedback Feedback	Math overall L L L L L L L L L	L L L L L L L L L
genderMatching Gender matching	Education enrollment L L L L L L L L L	L L L L L L L L L
planT Plan T (camp)	Education enrollment L L L L L L L L L	L L L L L L L L L

References on included studies

- Andersen, S., Beuchert, L., Nielsen, H., & Thomsen, M. (2018). The effect of teacher's aides in the classroom: Evidence from a randomized trial. *Journal of the European Economic Association*, 18. <https://doi.org/10.1093/jeea/jvy048>
- Andersen, S., Guul, T., & Humlum, M. (2019, August). *Instruction time, information, and student achievement: Evidence from a field experiment* (Working paper).
- Andersen, S. C., Guul, T. S., & Humlum, M. K. (2022). How first-language instruction transfers to majority-language skills. *Nature Human Behaviour*, 6(2), 229–235.
- Andersen, S. C., Humlum, M. K., & Nandrup, A. B. (2016). Increasing instruction time in school does increase learning. *Proceedings of the National Academy of Sciences*, 113(27), 7481–7484. Retrieved from <https://doi.org/10.1073/pnas.1516686113>
- Andersen, S. C., & Nielsen, H. S. (2016). Reading intervention with a growth mindset approach improves children's skills. *Proceedings of the National Academy of Sciences of the United States of America*, 113(43), 12111–12113. Retrieved from <https://www.jstor.org/stable/26472210>
- Andersen, S., & Nielsen, H. (2020). Learning from performance information. *Journal of Public Administration Research and Theory*, 30(3), 415–431. Retrieved from <https://doi.org/10.1093/jopart/muz036>
- Beuchert, L., Eriksen, T. L. M., & Krægpøth, M. V. (2020). The impact of standardized test feedback in math: Exploiting a natural experiment in 3rd grade. *Economics of Education Review*, 77, 102017. Retrieved from <https://doi.org/10.1016/j.econedurev.2020.102017>
- Bleses, D., Højen, A., Dale, P. S., Justice, L. M., Dybdal, L., Piasta, S., Markussen-Brown, J., Kjærbaek, L., & Haghighi, E. (2018a). Effective language and literacy instruction: Evaluating the importance of scripting and group size components. *Early Childhood Research Quarterly*, 42, 256–269. Retrieved from <https://doi.org/10.1016/j.ecresq.2017.10.002>
- Bleses, D., Højen, A., Justice, L., Dale, P., Dybdal, L., Piasta, S., Markussen-Brown, J., Clausen, M., & Ebad Fardzadeh, H. (2018b). The effectiveness of a large-scale early childhood language and pre-literacy intervention: The SPELL randomized controlled trial in Denmark. *Child Development*, 89(4), e342–e363. Retrieved from <https://doi.org/10.1111/cdev.12859>
- Bleses, D., Højen, A., Kjær Fauriskov, M., Line, D., & Sehested, K. (2015, April). *Sproget kan styrkes!: Rapport om effekten af sprogindsatserne SPELL og Fart på sproget i danske dagtilbud*. Syddansk Universitet.

- Bleses, D., Jensen, P., Højen, A., Slot, P., & Justice, L. (2021). Implementing toddler interventions at scale: The case of "We learn together." *Early Childhood Research Quarterly*, 57, 12–26. Retrieved from <https://doi.org/10.1016/j.ecresq.2021.04.008>
- Bleses, D., Jensen, P., Slot, P., & Justice, L. (2020). Low-cost teacher-implemented intervention improves toddlers' language and math skills. *Early Childhood Research Quarterly*, 53, 64–76. Retrieved from <https://doi.org/10.1016/j.ecresq.2020.03.001>
- Bleses, D., Klyvø, L., Willemsen, M., & Boisen, L. (2022, February). *Hovedresultater fra forskningsprojektet READ i plejefamilier*. TrygFondens Børneforskningscenter.
- Bleses, D., Moos, M., & Boisen, L. A. (2020). *Hovedresultater fra forskningsprojektet VLS+*. TrygFondens Børneforskningscenter.
- Bleses, D., Trecca, F., Højen, A., Justice, L., Slot, P., & Purtell, K. (2025). Effects of an infant/toddler intervention at 1-year follow-up: Sustained impacts to preschool entry. *Educational Researcher*, 54(1), 21–33.
- Bleses, D., Vind, B., & Jensen, A. (2024, January). *Effekter af læseleg i vuggestuen og dagplejen*.
- Damm, A. P., Mattana, E., Nielsen, H. S., & Rouland, B. (2021). Academic achievement and wellbeing of dual language learners: Evidence from a busing program. *Journal of Urban Economics*, 126, 103358. Retrieved from <https://doi.org/10.1016/j.jue.2021.103358>
- Gissel, S. T., & Andersen, S. C. (2021). A cluster-randomized trial measuring the effects of a digital learning tool supporting decoding and reading for meaning in grade 2. *Journal of Computer Assisted Learning*, 37(2), 287–304.
- Guul, T. S. (2018). The individual-level effect of gender matching in representative bureaucracy. *Public Administration Review*, 78(3), 398–408. Retrieved from <https://www.jstor.org/stable/26648896>
- Harrer, M., Cuijpers, P., Furukawa, T. A., & Ebert, D. D. (2021). *Doing meta-analysis with R: A hands-on guide* (1st ed.). Chapman & Hall/CRC Press.
- Hvidman, C., Koch, A., Nafziger, J., Nielsen, S. A., & Rosholm, M. (2020). *An intensive school-based learning camp targeting academic and non-cognitive skills evaluated in a randomized trial* (Technical report). IZA – Institute of Labor Economics. Retrieved from <https://www.jstor.org/stable/resrep61202>
- Jakobsen, M., & Andersen, S. C. (2013). Coproduction and equity in public service delivery. *Public Administration Review*, 73(5), 704–713. <https://www.jstor.org/stable/42002998>

Nandrup, A. B. (2016). Do class size effects differ across grades? *Education Economics*, 24(1), 83–95. Retrieved from <https://doi.org/10.1080/09645292.2015.1099616>

Nielsen, S. A. (2023). Boys left behind: The effects of summer camp and follow-up strategies on academic, personal, and social competencies. *Economics of Education Review*, 93, 102370. <https://doi.org/10.1016/j.econedurev.2023.102370>

Rosholm, M., Hvidman, C., Nielsen, S., Gumede, K., Nafziger, J., & Koch, A. (2020, March). *Turboforløb til ikke-uddannelsesparate elever i 8. klasse: Afsluttende rapport*. Trygfondens Børneforskningscenter.

Rosholm, M., Mikkelsen, M. B., & Gumede, K. (2017). Your move: The effect of chess on mathematics test scores. *PLoS ONE*, 12(5), e0177257.

Rosholm, M., Mikkelsen, M. B., & Svarer, M. (2019). Bridging the gap from welfare to education: Propensity score matching evaluation of a bridging intervention. *PLoS ONE*, 14(5), 1–20. <https://doi.org/10.1371/journal.pone.0216200>

Rosholm, M., Nielsen, S. A., Hvidman, C., et al. (2020). *Effektevaluering af tre intensive læringsforløb*. TrygFondens Børneforskningscenter.

Rosholm, M., Tonnesen, P. B., Rasmussen, K., Overgaard, S., Færch, J. V., Malm, S. G., & Harder, J. (2025). A tailored small group instruction intervention in mathematics benefits low achievers. *npj Science of Learning*, 10(1), 18. Retrieved from <https://doi.org/10.1038/s41539-025-00310-9>

Rosholm, M., Vind, B. D., Laursen, N. B., Lomholt, J. J., & Nielsen, S. A. (2019). *Evaluering af skolerettede og skoleunderstøttende indsatser i projekt "Styrket faglighed blandt udsatte børn i folkeskolen"*.

Svarer, M., Rosholm, M., Havn, L., & Høeberg, L. (2014). *Evaluering af mentorindsats til unge uden uddannelse og job*. Rambøll.

Comments on Stine Nyhus Larsen et al.: Childhood and Adolescent Interventions in a Danish Setting

Marjo Kyllönen and Kristiina Hannukainen

The Larsen et al. article aimed to estimate pooled effects on children's learning and to explain between-study variation using cost-standardised effect sizes and coded information on intervention content, delivery, and context collected during a principal investigator survey. The findings indicate that education interventions in Denmark have, on average, led to minor improvements, with meaningful variations in the effects and cost-effectiveness due to the design and context.

The effects of educational interventions, particularly in the context of the Nordic welfare state – specifically in Finland – are of significant interest to us. As noted in the article, the focus on interventions also serves as a research tool to highlight the inequalities in income and wealth that affect learning and wellbeing.

It was a pleasure to receive an article for evaluation that is timely and addresses a topic that has not been sufficiently scrutinised from the perspective of education providers. This evaluation also affords us, as educational organisers, the opportunity to articulate our insights regarding the significance of collaboration between academia and practice in shaping urban policy.

1 The City of Helsinki relies on research data to develop its operations

The article by Larsen et al. was interesting and inspiring. We found many similarities with the areas we are also eager to explore and understand better in Helsinki. The City of Helsinki is the largest education provider in Finland, serving approximately 200,000 individuals, from early childhood to adult education. We have over 15,000 professionals working at more than 700 locations. Our goal is to foster positive learning and wellbeing for all citizens in Helsinki while working collaboratively to build a sustainable future.

We need to work towards equitable learning, strengthen core knowledge and skills, and support the development of social and emotional competencies. In our Education 2030

strategy (City of Helsinki, 2024), we emphasise the importance of research and data. We recognise that to succeed in our mission, our efforts must be guided by research and evidence. As such, we leverage research data to promote growth and development in our learners.

We encourage our schools and educators not only to adopt a variety of teaching methods but also to use our research data to assess and enhance pedagogical resources. As highlighted in the article, high-quality interactions between teachers and students, well-targeted curricula, enriched learning environments, and support for teachers' professional development are linked to positive educational outcomes.

2 Why is collaboration between academia and decision makers essential?

The Mayor of Helsinki recently unveiled a new city strategy for 2025-2029 (City of Helsinki, 2025), stating, "We rely on research data in the development of our operations." While we undertake our own research, we firmly believe that collaboration with local universities is vital for enhancing and developing our initiatives. Helsinki aspires to be a pioneering force in impact research and its application for the betterment of children and young people.

We recognise that education constitutes a significant investment in the future, with half of the city budget dedicated to educational services. We must ensure our resources are allocated and utilised effectively. For example, we are channelling needs-based funding into early childhood education and care (ECEC) and schools in less affluent socio-economic districts. In Helsinki, our goal is to alleviate the adverse effects of regional segregation on educational institutions and learners through targeted funding in the education sector. A critical question arises: how can we ascertain that utilising these resources effectively reduces disparities?

The phenomena of polarisation and segregation are escalating challenges in Helsinki. While our educational ethos emphasises principles of equality and equity, the latest PISA research (2022) indicates that basic education is falling short in bridging the gaps.

As educational providers, our challenge is to determine whether our structures and pedagogical practices have a positive impact on children's growth and learning outcomes. Randomised controlled trials (RCTs) are integral to our broader evaluation framework. These experiments furnish education providers and policymakers with valuable insights on enhancing educational practices and improving learning outcomes.

The findings presented by Larsen and colleagues have significant implications for Helsinki: interventions aimed at teachers in schools and educators in daycare and

preschool settings are markedly more effective than those directed at children or their parents. Fully manualised interventions tend also to be less effective than non-manualised or partially manualised ones. Results show that language outcomes typically exhibit substantially lower effect sizes than those in mathematics, whereas other outcomes, often related to vocabulary acquisition, demonstrate larger effects. Interventions designed for children under school age (preschoolers) are more effective than those targeting older children. Notably, while interventions that prioritise quality enhancement generally yield lower effects than those focused on increasing quantity, those emphasising structural quality tend to be more effective than those centred on process quality or learning formats.

Interestingly, the evaluation design (RCT versus other methodologies) does not significantly influence effect sizes. This indicates that, on average, non-experimental evaluation designs do not substantially overestimate or underestimate true effect sizes (as determined by RCTs) in our sample. However, this does not imply the absence of bias in individual studies, as different forms of bias may take various forms and potentially offset each other on average.

Language assessments yield significantly poorer outcomes compared to mathematics and other measures. Interventions aimed at preschool children are more cost-effective than those focused on older children and young adults, especially when delivered in daycare settings. Additionally, interventions targeting educators are more cost-effective, whereas fully manualised interventions tend to be less so.

To achieve the objectives stipulated in our City Strategy—that every local school in Helsinki is a high-quality educational institution—we require a more extensive, research-informed understanding of the effectiveness of routine practices in schools and early childhood education.

To ensure that every child and young person acquires the necessary skills and competencies to succeed in further education and professional life, we must deepen our understanding of how best to support our schools and teachers in choosing the most effective pedagogical approaches. It is important to note that in Finland, teachers have considerable pedagogical autonomy in selecting classroom tools and methods, a freedom we respect and value.

We are particularly eager to deepen our understanding of how to improve the learning experiences and developmental outcomes of disadvantaged children and young people.

3 Conclusions

This article and its meta-analysis present intriguing findings that resonate with our current focus in research collaboration with academia. Here are some reflections based on the article that warrant consideration:

- Carefully designed interventions have a positive impact on children's learning outcomes at all ages.
- Interventions aimed at younger children are particularly effective and promising.

One critical observation from the meta-analysis is that interventions targeting educators have been more effective than those directed at children and their parents.

- Given that a pupil's background is the strongest predictor of learning outcomes, how can we identify effective methods to support and guide families?

In conclusion, this aligns with our vision outlined in the new City Strategy: a Helsinki of which we can be proud (City of Helsinki, 2025). Realising our vision will require research partnerships with scholars and universities, and the insights in the article by Larsen and colleagues are particularly relevant to our new strategy. We need research collaboration with universities, as well as long-term, well-designed interventions that also target preschool and early childhood education.

References

City of Helsinki. (2024). *Helsinki learns – Future competences: Education 2030*. Retrieved from: <https://www.hel.fi/static/kasko/helsinkioppi/helsinki-learns-future-competences.pdf>

City of Helsinki. (2025). *Helsinki city strategy 2025–2029*. Retrieved from: <https://www.hel.fi/en/decision-making/strategy-and-economy/strategy>

Comments on Stine Nyhus Larsen et al.: Childhood and Adolescent Interventions in a Danish Setting

Hannu Karhunen

Building on an earlier survey by Rosholm et al. (2021), Stine Larsen et al. present an updated meta-analysis of interventions targeting children conducted in Denmark in collaboration with TrygFonden's Centre for Child Research. Given that children and adolescents constitute one of the most vulnerable groups in society, interventions aimed at improving their learning and wellbeing require particularly careful and rigorous evaluation. As the authors point out, randomised controlled trials (RCTs) are considered the gold standard in causal inference. However, quasi-experimental methods provide a valuable complement for estimating causal effects, particularly when RCTs are not feasible.

Extending the earlier analysis from purely experimental studies to include quasi-experimental evaluations is one of the article's main contributions, alongside the provision of cost-standardised effect size estimates. The findings can be summarised as follows. First, RCTs and non-RCTs yield similar effect sizes, suggesting that quasi-experimental estimates are, on average, not systematically biased. Second, the authors find support for the Heckman curve hypothesis, which holds that estimated effect sizes of interventions tend to be larger when they are implemented earlier in life.

Third, and perhaps the most surprising finding of the paper, the authors show that the costs of interventions do not correlate with the effect sizes. Combined with the result that interventions focusing on quality improvements appear to be less effective, this raises important questions about the Nordic model, which invests heavily in children's education and wellbeing but may not always achieve commensurate outcomes. As discussed earlier by Rosholm et al. (2021), the poor quality of educational interventions may help explain why substantial investments by Nordic welfare states do not translate into greater improvements in equality of opportunity.

Which interventions seem to be the most cost-effective? Overall, the two most cost-effective interventions are both quasi-experimental and rely on naturally occurring institutional features, showing that low-cost informational and feedback mechanisms can have substantial educational impact relative to their costs. The most cost-effective intervention is the effect of receiving national test results in mathematics and reading.

This quasi-experimental study exploits an unplanned IT breakdown during the introduction of national testing in Denmark, which created random variation in whether students received information about their current performance (see Andersen and Nielsen, 2020). This design allowed the authors to identify the causal impact of performance information itself, showing that simply informing students and teachers about achievement levels can significantly improve subsequent performance at very low cost.

The second most cost-effective intervention concerns feedback based on performance categories, in which small differences in test scores lead to different evaluative labels (see Beuchert et al., 2020). This study also uses a quasi-experimental design based on the mechanical conversion of test scores into five performances. The study examines how this relatively negative feedback affects subsequent achievement, demonstrating substantial effects at minimal cost.

While the article focuses solely on Danish studies conducted in collaboration with TrygFonden's Centre for Child Research, it provides broader insights that should be considered in all Nordic countries. The results highlight the need to prioritise rigorous evaluation, using both experimental and quasi-experimental methods when designing and scaling up interventions or actual policies. Evidence of many insignificant, and even some negative, effects suggests that well-meaning interventions can, at times, be counterproductive for young people. Extensive regional or school-level autonomy, common in Nordic countries, should not be used as a reason to forgo systematic impact evaluation.

The Larsen study, along with other recent evidence – particularly from Norway (e.g., Haaland et al., 2024) – also demonstrates how well-functioning research infrastructure and close collaboration with local and governmental authorities are crucial for generating robust, research-based evidence on early-life interventions. For example, many of the studies surveyed by Stine Larsen et al. would not have been possible without extensive infrastructure, such as national standardised assessments of student learning at different stages of schooling in Denmark. Comparable infrastructure exists in all Nordic countries – except Finland – thereby enabling evidence-based development and evaluation of early-life interventions. However, it is still too early to assess the extent to which this infrastructure is being fully utilised.

References

- Andersen, S., & Nielsen, H. (2020). Learning from performance information. *Journal of Public Administration Research and Theory*, 30(3), 415–431.
- Beuchert, L., Eriksen, T. L. M., & Krægpøth, M. V. (2020). The impact of standardized test feedback in math: Exploiting a natural experiment in 3rd grade. *Economics of Education Review*, 77, 102017.
- Haaland, V. F., Rege, M., & Solheim, O. J. (2024). Do students learn more with an additional teacher in the classroom? Evidence from a field experiment. *The Economic Journal*, 134(657), 418–435.
- Rosholm, M., Paul, A., Bleses, D., Højen, A., Dale, P. S., Jensen, P., Justice, L. M., Svarer, M., & Andersen, S. C. (2021). Are impacts of early interventions in the Scandinavian welfare state consistent with a Heckman curve? A meta-analysis. *Journal of Economic Surveys*, 35(1), 106–140.



Nordic Economic Policy Review 2026

Large-Scale Randomised Experiments with Population Data

Elias Einiö and Annika Nivala

Abstract

This article features insights from the Finnish Recruitment Subsidy Experiment (Einiö et al., 2021; Einiö and Nivala, 2026), which we designed and implemented along with the Ministry of Economic Affairs and Employment of Finland. It belongs to a class of randomised controlled trials characterised by large, national-scale, costly treatments and imperfect take-up, which are typical features of many public policies. The experimental setting is based on data drawn from population registers covering all individuals, firms or other relevant units of observation. Data of this sort are becoming increasingly common in many countries, and are particularly accessible in the Nordic Region. We address econometric issues typical for these types of settings, including the management of over- and under-subscription, as well as the calculation of statistical power with multiple experimental waves. We also discuss practical issues relevant to running large-scale national randomised experiments involving a research team and several parties from the public administration.

JEL codes: H25, J38, L26, M51

1 Introduction

Randomised experiments are a common tool for estimating the causal impacts of policy interventions, yet there is great variation in terms of experimental settings. One factor that often limits the design of experiments is the cost of data collection. When data needs to be collected in multiple survey waves, both before and after the experiment, the costs increase with the number of waves and observations. In this article, we consider a specific class of experiments in which such limitations are minimal – namely, randomised controlled trials conducted in settings where data are available in population registers covering all individuals, firms or other relevant units of observation.

We address econometric topics that are relevant to large-scale experiments with register data, and offer insights from an experiment that we designed and implemented in conjunction with the Ministry of Economic Affairs and Employment of Finland: the Finnish Recruitment Subsidy Experiment (Einiö et al., 2021; Einiö and Nivala, 2026). The article is not intended to be a comprehensive guide to designing randomised experiments, but rather a discussion of specific issues that may arise in the implementation of large-scale randomised experiments that use register data and involve imperfect take-up and costly treatments (e.g., monetary subsidies). These are commonly occurring features in public support policies, whether they target individuals or firms, and arise in the context of national experiments on such instruments, conducted with the public administration.

The availability of population register data has several advantages, an important one being the elimination of data-collection costs.^[26] Population data also enable large-scale experiments that facilitate highly accurate estimates of treatment effects. The ready availability of data also helps with planning and implementing experiments. Importantly, population register data allows for an estimation of the effects of offering a treatment, which is essential in many public policy contexts in which individuals are not required to take up a subsidy, transfer or other form of public support.

In the economics literature, the effect of offering a treatment is often referred to as the intention-to-treat (ITT) effect. It is widely viewed as the single most important parameter for policy-making. The difference between ITT and the treatment effect on the treated units – which is often the focus and the relevant parameter, e.g., in medical studies – is that some units do not necessarily take up the treatment. For instance, a firm may not take up an investment subsidy if the subsidy does not make new investments profitable enough, or an individual may not take up a student allowance if

26. Such data must be gathered by the administration, of course, but it is primarily collected for other purposes, such as taxation, and therefore need not be counted as a research cost.

it does not make studying the most beneficial long-term option.^[27] More generally, in many public policy settings, public expenditure on a unit – be it an individual or a firm – is determined by both eligibility and the unit's own actions.

The importance of the intention-to-treat parameter increases with the share of units that do not take up the treatment, making it a key parameter of interest in many public policy contexts (e.g., Duflo et al., 2007). By estimating the ITT effect, it becomes possible to predict the effects of a policy when it is made available to a wider population – this effect arises from both the impact of the treatment on the treated and the proportion of individuals who take it up. For instance, an investment subsidy that doubles investment in a firm that takes up the treatment has little aggregate impact if only a few firms make use of it (e.g., if the subsidised investment is deemed insufficiently profitable). The ITT estimate recovers the average effect of offering a treatment at population level, which is often the primary concern for decision-making.

Experiments are almost always limited by budget constraints.^[28] It may be desirable, for multiple reasons, to use the whole of the budget. When conducting experiments with the public administration, the budget may be set as part of a political process, with the intention of spending it all. When the maximum number of treatments are distributed, the experiment's ability to detect treatment effects with sufficient statistical precision typically improves compared to when a smaller number of treatments are distributed. In this article, we discuss the potential challenges that a fixed budget may introduce in settings where take-up is not perfect – a feature of many public policies – and not known in advance, along with strategies for managing over- and under-subscription to the treatment.

The second part of the article features lessons from the Finnish Recruitment Subsidy Experiment, which we conducted with the Finnish Ministry of Economic Affairs and Employment. The goal of the experiment was to evaluate whether such a subsidy is effective at inducing non-employer entrepreneurs to hire their first employees and to what extent this would spur growth among these firms. We discuss issues that are relevant in practice when running large-scale national randomised experiments involving a research team and several parties from the public administration. We describe the planning, timeline, implementation and evaluation of the experiment, as well as data management issues. We briefly discuss the results of the experiment and compare them to a previous observational study on a similar policy instrument. The final part of the article provides recommendations and concluding remarks.

27. By contrast, in medical studies, the interest is in the effect of the substance, which means that full take-up is important for a successful experiment. However, the aggregate effects of offering the substance to a wider public will depend on many other factors, including the prevalence of conditions cured by it, and the willingness of individuals who suffer from the conditions concerned to take care of themselves.

28. Information treatments are an exception, as the costs of sharing information to an additional treated unit are negligible. However, even information treatments can have some marginal cost, e.g. if the information needs to be sent in hard copy by post.

2 Experiments with imperfect take-up and population data

In this section, we discuss some econometric issues specific to large-scale experiments with population data and imperfect take-up. Our practical example is the recruitment subsidy for first hires, which is conditional on becoming an employer or incurring positive wage costs. More generally, the setting we discuss corresponds to a randomised experiment in which the treatment group is offered a treatment (e.g., a monetary subsidy) that is conditional on taking an intended action. The agent is not forced to take the action, but the treatment offer increases their incentives to choose it. This is a typical treatment setting for many public policies, such as investment subsidies for firms or allowances for students. The control group may also choose to take the intended action (e.g., invest or study), even without being offered the additional incentive. Other examples of such policies are free vaccines, bonus pension coverage from additional years of work, and so forth.

The goal of the experiment is to assess the effect of the treatment not only on the likelihood of choosing the intended action, but also on other outcomes, whether intended or not.

2.1 Under- and over-subscription

Experiments are rarely performed without budget constraints. The budget, including the cost per treatment, determines the number of units to which the treatment can be granted. When the take-up rate is uncertain, two possible situations arise: (i) under-subscription, where fewer units take the treatment than allowed for in the budget; and (ii) over-subscription, where more units would take the treatment than it is possible to accommodate.

Suppose the researcher estimates the take-up rate at 5% – that is, 5% of firms in the treatment group would take up the subsidy. In addition, suppose that the budget can cover 10,000 treatments (e.g., a monetary subsidy). How many firms should be offered the treatment? This decision would be an easy one if there were no uncertainty about the take-up rate. However, because the rate needs to be estimated, there is a possibility that the actual take-up rate will be larger than the estimate and too many will accept the treatment (over-subscription); or the take-up rate could be lower, leading to fewer accepting the treatment than the budget can accommodate (under-subscription). Therefore, if the researcher prefers under-subscription to over-subscription, they should offer the subsidy to fewer firms than the estimate of the take-up rate would imply, and vice versa.

Whether a researcher prefers over- or under-subscription depends on the weight they give to problems arising from each of these scenarios. Under-subscription means that more units could have been offered the treatment, given the actual take-up rate. While this does not bias the experimental results, a smaller treatment group size reduces the statistical power of the experiment. Conversely, over-subscription means that some firms that would like to accept the treatment will not receive it. If those firms are somehow selected, this may lead to bias in the estimate of the treatment effect. We discuss the potential biases introduced by over-subscription in Section 2.4.

In light of the previous discussion, over-subscription is often a relatively bigger concern than under-subscription, as the former may lead to bias in the estimation. However, significant under-subscription is also problematic because it may lead to underpowered statistical testing, making the experiment less useful. As such, it is important to balance the trade-offs between over- and under-subscription.

2.2 Managing over-subscription: A multi-wave setting

In this section, we briefly discuss an approach to recovering the take-up rate to calculate the treatment group size – namely, using a waved experiment comprising at least two experimental waves. The aim of the first wave is to estimate the take-up rate. In this wave, the number of treatment offers is large enough to provide a sufficiently reliable estimate of the rate, but small enough that the risk of over-subscription is negligible. The minimum number of waves for this approach is two, but the larger the number, the more accurate the estimate of the targeted number of treatments. The trade-off, however, is that the duration of the experiment increases with the number of waves.

Once the take-up rate has been estimated in the first wave, the size of the second wave can be determined, based on a statistical calculation that takes into account the uncertainty of the first-wave take-up rate estimate. The aim is to avoid over-subscription without significantly reducing the experiment's size and power. In this approach, the researcher decides the maximum risk of over-subscription that they are willing to accept. They can then calculate an adjusted value for the take-up rate that produces over-subscription with that probability. We discuss the details of this approach in the Technical Appendix.

Another approach that does not rely on estimating the take-up rate is to run multiple waves with small treatment and control group sizes, until the desired number of allocated treatments has been achieved. The practical problem with this approach is often time: in many policy interventions, individuals need a reasonably long time to respond and take up the treatment. Such an experiment would therefore take a long time to run, unless the response time can be very short. What is considered a

reasonable time to respond depends, of course, on the type of treatment. In our example of firms considering hiring their first employee, planning and acquiring the required information may take some time.

A third approach is to randomise the applicant pool. While this type of design can identify the treatment effect of receiving the subsidy, it is not necessarily informative about population-level effects, because the procedure leads to a selected experimental population. For instance, some firms that would use the subsidy on offer to hire employees might not apply for it if receiving it is a lottery rather than a certainty. This type of selection might be particularly important for entrepreneurs at the margin of employer entry, whose response is an important component of the ITT effect.

2.3 Power in a multi-wave setting

When pre-planning RCTs, it is essential to calculate the experiment's statistical precision. These calculations inform the planner of the degree of precision in the estimates of the expected treatment effects. The key factors that determine statistical power are the variation in the outcome variables, the size of treatment and control groups, and the level of randomisation.

For several types of single-wave RCTs, mathematical formulas for statistical power are available, and the calculations are straightforward (see, e.g., McConnell and Vera-Hernandez, 2015). For more complex settings, formulas may not be available. In such cases, Monte Carlo methods can be used to simulate the statistical power. In addition, simulations can be used to evaluate the magnitude and likelihood of over-subscription bias and are a useful tool for developing sampling algorithms.

The size of the target population can be a limiting factor for large-scale randomised experiments with a low take-up rate. If the take-up rate is low, very large treatment groups may be required in order to reach the desired number of units taking up the treatment. In some cases, the treatment group could be so large that it accounts for more than half of the target population, while the number of units left for the control group could be small. If the control group is too small, the precision of the experiment is reduced. In that case, rather than aiming to allocate as many subsidies as possible, the most effective design can be determined via methods that identify the optimal split into treatment and control groups (e.g., McConnell and Vera-Hernandez, 2015).

2.4 Selection of treated units due to over-subscription

In this section, before proceeding to the practical example of the Finnish Recruitment Subsidy Experiment, we discuss the biases associated with over-subscription.

Over-subscription means that the treatment cannot be given to all units in the treatment group that want it. If those units that would have accepted but did not receive the treatment are systematically selected, this can introduce bias into the estimates. This can be a particular concern in settings in which the public administration grants treatment to treatment group units based solely on the order in which they applied, as is typical in public policy settings. For example, in a setting where the public administration stops granting treatments when the maximum number has been allocated, those who apply last will be left out. If those units had only taken the intended action (e.g., invest or study) after receiving the treatment, their exclusion from the group of units treated can lead to bias in the ITT estimate of the impact of offering the treatment on the incidence of taking the intended action, moving it in a downward direction.

3 Lessons from practice: The Finnish Recruitment Subsidy Experiment

The Ministry of Economic Affairs and Employment commissioned the Finnish recruitment subsidy experiment in 2022–2023. The experiment followed the approach described above: a subsidy was offered to the treatment group with ex ante uncertain, imperfect take-up, within a budget that allowed a maximum of 900 subsidies. The experiment was run in two waves, and the treatment group size was adjusted based on the take-up rate in the first wave.

All of the data needed to plan, implement and evaluate the experiment was pre-existing register data, with the exception of the Experiment Register, that was specifically created to run the experiment and administer subsidy payments. There was no need for extensive data collection outside of normal administrative processes. The key public register that enabled the experiment was the Incomes Register, which requires all employers to report wages paid to employees on a monthly basis.

3.1 Motivation

The economic motivation for the first employee subsidy stems from the idea that becoming an employer incurs some fixed costs or market frictions that hamper a firm's growth. For example, a non-employer firm needs to learn about employer responsibilities and recruitment before hiring, and there may be uncertainty about the productivity of the first employees. These considerations lead to uncertainty about the profitability of hiring and costs before the firm receives any potential profits from the new employees' input. These types of costs, including time spent on hiring and

acquiring information on employer liabilities, can create barriers to becoming an employer. In such cases, a temporary subsidy for hiring could incentivise a firm to become a permanent employer and accelerate its long-term growth.

There is little evidence regarding the extent to which these factors limit firm growth. However, over half of all firms have no employees, indicating a large potential target group for a recruitment subsidy.

3.2 Planning process and timeline

A recruitment subsidy experiment aimed at encouraging hiring by non-employer firms was part of the 2019–2023 government programme. The Ministry began planning the experiment in early 2020 and commissioned us to design it (Einiö et al., 2021). The planning process took place in summer 2020. The Ministry then prepared to conduct the experiment and drew up a legal proposal. During this time, we did additional consulting work on running the experiment, including statistical power calculations based on register data, technical feedback on the proposed legislation, and drafting the letters sent to the treatment group of firms. Parliament passed the legislation on the Recruitment Subsidy Experiment in January 2022 (FINLEX, 2022).

As part of the planning process, we drew up an idealised proposal for the experiment. The Ministry took the time to respond to this proposal and set constraints based on feasibility. The final plan incorporated these feasibility constraints as well as suggestions by policy-makers. The plan struck a balance between the ideal experiment, which tested a recruitment subsidy to increase first hires, and administrative and political feasibility. The experiment and subsidy design were close to our initial proposal, but there were some changes, as discussed below.

We argued for randomisation at the level of the whole target population, to be identified by the tax authority on the basis of register data. This allowed us to estimate a population-wide intention-to-treat effect, the key parameter policy-makers use to decide whether to implement the policy on a larger scale. Another option discussed was randomisation from the applicant pool. As discussed in Section 2.2, while this type of design can identify the treatment effect of receiving the subsidy among the applicants, this treatment effect in itself is not necessarily informative about population-level effects.

When identifying the effects of offering subsidies, one issue to consider is whether they have equilibrium effects. These can lead to spillover effects on control firms, or other firms, which can make it difficult to disentangle the treatment effect. In addition, even if the effects of a smaller-scale pilot experiment are clearly identified, they may not reflect the effects of adopting the policy at full scale. The reason is that equilibrium

effects may be more likely to arise in a full-scale policy due to a larger number of treated units. In the recruitment subsidy experiment, the subsidy cannot be expected to have significant equilibrium effects, at least in the short run, because first hires by non-employer firms account for a fairly small proportion of the labour market.

One key aspect of the subsidy design was to minimise application costs and bureaucracy for the entrepreneur. An increased administrative burden may lead firms to decide not to apply for subsidies, which can be a major reason for their ineffectiveness as an incentive. The experiment therefore made it as easy as possible for firms to apply for and be granted the subsidy. This both maximised effectiveness and ensured that the results reflect how well the subsidy reduces the barriers to hiring first employees, rather than frictions in take-up due to administrative complexity and costs.

We suggested that receiving the subsidy would be automatic, based on the firm's reported wage costs in the Incomes Register, without the need for an application process. However, this was not feasible due to business subsidy regulations, which require firms to provide certain information in order to be granted a subsidy, including adherence to the EU de minimis rule. As a result, the firms in the treatment group that wanted the subsidy had to apply for it via a single, short online application. After this initial application, the subsidy was paid automatically based on the wage costs in the Incomes Register. We also suggested ruling out the hiring of family members with the subsidy, to reduce the possibility of misuse. However, this was not considered administratively feasible.

The experiment required a time limit for subsidy use. The initially suggested timeframe for hiring was one year, but it ended up being shorter, at five months, which may not have been long enough for some firms.

The experiment was facilitated by access to register data at all stages: planning, implementation and evaluation. We used firm tax return data from previous years to aid in planning the experiment, including selecting the target group. The target group was identified from the registries, and randomly allocated to treatment and control groups. Automatic subsidy payments (conditional on the initial application) were based on the Incomes Register, which allowed real-time payments to be made with very little lag with respect to realised wage costs and administrative costs for firms. Due to the ready availability of the register data, there was no need for extensive data collection, which enabled us to evaluate the effects quickly and at relatively low cost.

3.3 Use of previous evidence in planning

The planning of experiments can benefit substantially from previous research on similar instruments and outcomes. Previous research can inform the researcher both about the distributions of outcomes and the expected take-up rates and effects of the instrument. This information can be very useful even if it shows no significant impacts, as it can help researchers design more effective treatments. Overall, previous evidence can help to inform both the design of the treatment and the details of the experiment.

For the recruitment subsidy experiment, one source of information that benefitted the planning was an earlier regional support instrument that encouraged non-employer firms to hire their first employees. Nivala (2024) studied the effects of this previous regional subsidy programme using a difference-in-differences design, which compared regions in which firms were eligible for the subsidy to regions in which similar firms were ineligible. In Section 3.10, we compare the results from the observational evaluation of the regional subsidy and the results from the RCT evaluation of the Recruitment Subsidy Experiment (Einiö and Nivala, 2026). Here we discuss how the experiences from the regional experiment informed the planning of the Recruitment Subsidy Experiment.

The previous regional first-employee subsidy programme was implemented in parts of Finland in 2007–2011. It targeted firms that had no employees for at least 12 months. To qualify, a firm had to hire an employee on a permanent full-time contract (defined as 25 hours per week). The subsidy covered 30% of wage costs in the first year and 15% in the second year. Firms had to apply for the subsidy before hiring and then apply for subsidy payments, which were paid out twice a year.

Nivala (2024) finds that the programme had no effect on the probability of hiring. A striking observation is that the regional subsidy had an extremely low take-up rate: only 2% among new employers. Low take-up rates can be due to scheme complexity, poor salience of the programme or high application costs (e.g., Bhargava and Manoli, 2015; Finkelstein and Notowidigdo, 2019). One key detail that may have discouraged firms from using the subsidy was the requirement to hire on a full-time, permanent contract. Additionally, application costs and payment delays may have reduced the desirability of taking up the subsidy.

While this prior evidence does not provide good estimates for a potential take-up rate in the RCT, it led us to think particularly carefully about which design features would improve take-up. It also helped convince the administration of the importance of introducing such features. As a result, significant efforts were made to ensure that the subsidy instrument was as simple and easy as possible. In the RCT, the subsidy was paid automatically, based on wage costs in the Incomes Register. In addition, letters

were sent directly to the treatment group, notifying them that they were eligible for a subsidy.

3.4 The recruitment subsidy

In the randomised experiment, the recruitment subsidy was 50% of the firm's total wage costs for the first 12 months, up to a maximum of €10,000. The target group included firms that: 1) had no employees (except the entrepreneur(s) themselves) in the previous 12 months; 2) had revenue of €15,000–1M in the previous year; and 3) had no tax debt, no ongoing bankruptcy filings, etc. The target group was identified based on tax registry data. The revenue constraint was introduced to increase the experiment's statistical power: excluding extremely small and large firms reduced the standard errors of outcome variables, which allowed for greater precision in the estimates. The third restriction was required by other business subsidy regulations.

Each of the firms in the treatment group was sent a letter notifying them of their eligibility for the subsidy. To qualify, they had to hire their first employee within four months after the month of receiving the letter. The letters included an explanation of the subsidy, information on how to apply and referrals to additional information.

To apply, firms had to complete a simple electronic application form administered by TE Services of Southeast Finland (hereinafter TE Services), a public employment services office under the Ministry of Economic Affairs and Employment, which provides business and employment services and allocates support for a range of programmes. The firms also had to provide the social security number of the hired employee, either on the application or as an addendum after applying. If a firm hired an employee before applying, the wage costs after the application were eligible for the subsidy. A firm could be denied the subsidy if it was not in the treatment group, or if its total of de minimis subsidies exceeded the allowed ceiling.

Once the subsidy was granted, it was paid automatically each month, based on the wage costs reported in the Incomes Register.

3.5 Randomisation

We conducted a stratified random sampling from the target population to treatment and control groups in two waves (see Section 3.6 for details). Stratification means that randomisation takes place within subgroups of units (called strata), which improves the balance between treatment and control groups. This approach enables the treatment effect of the subsidy programme to be estimated using a simple regression framework that compares the treatment and control groups within strata. The aim of

the randomisation is to ensure that the control and treatment groups are similar. The stratified sampling further balances the treatment and control groups based on the background characteristics that determine the strata. The randomisation was stratified by revenue quartiles within two-digit industries.

First, the tax administration identified the target population from tax registries and sent the list of firms and entrepreneurs to TE Services, which then ran our randomisation code written by us. When controlling for strata, the baseline characteristics of the treatment and control groups were found to be balanced – in other words, there were no statistically significant differences between the groups in terms of predetermined outcomes (Einiö and Nivala, 2026).

3.6 Waves

The experiment was conducted in two waves in order to manage the number of treatment-takers. The budget for the subsidy was fixed at €9 million, which meant 900 companies could be granted full subsidies. However, before the experiment began, the take-up rate in the target group was unknown. The fixed constraint on how many firms could be granted the subsidy, along with the ex ante uncertain take-up rate, raised a potential problem of over- or under-subscription, as discussed in Section 2.1. Conducting the experiment in two waves allowed us to manage the trade-off between these two.

The first, smaller wave allowed us to estimate the take-up rate. In the first wave, 3,500 firms were randomised into the treatment group, and 20,000 into the control group. The randomisation took place on 8 March 2022, letters to firms were sent by 14 March, and the firms had until the end of July to hire their first employee and apply for the subsidy.

The take-up rate in the first wave was estimated at 1.5%, which meant the subsidy could be offered to more than half of the target population. Hence, the sizes of the treatment and control groups in the second wave were set in such a way as to ensure an optimal split between them (e.g., McConnell and Vera-Hernandez, 2015). In the second wave, the treatment group size was set to 31,000 (rounded to the last hundred). In total, the treatment groups in both waves covered 34,500 firms, while the control groups in the rest of the target population totalled 38,771 firms.

The two-wave design had an important implication for the size of the treatment group. As a comparison, for a single-wave design, the statistical analysis and calculations in the original experimental plan provided a treatment group comprising 13,087 firms, which would have produced a significantly smaller number of treatment-takers and diminished the statistical power. This comparison demonstrates that wave design is a fundamental tool when running public policy experiments with incomplete take-up.

In the second wave, the randomisation was performed on 10 August 2022. At the beginning of August, the tax administration performed the same target group identification as in the first wave, but based on up-to-date data. The randomisation code excluded firms that were in the treatment or control groups in the first wave, and then randomised the rest of the target group. The letters were sent between 15 and 29 August. As the second-wave treatment group was substantially larger, it took somewhat longer for the TE Services to send out the letters.

3.7 Allocation of subsidies

TE Services administered the allocation of subsidies. This was done via a regional office that was responsible for Southeast Finland, but handled the treatment allocation throughout the whole country. The office was also responsible for communicating with the treatment group (e.g. the content of the letter informing the treatment group companies of their eligibility), which was planned together with the research team. This regional office was chosen because it had previously administered some temporary COVID subsidies, so it had experience of working on similar policies.

3.8 Data management

The authors, operating under the licence of their research institute (VATT), undertook the planning phase of the experiment based on access to firm tax data and accessed through Statistics Finland's secure online research data environment (FIONA). For the legal planning and randomisation phase, the tax authority provided VATT data on the target population, along with the variables needed to design the randomisation algorithm. TE Services were provided the wider dataset needed to meet its legal requirements and carry out the sampling based on the randomisation program provided by the researchers. This data also formed the basis for the Experiment Register, administered by the KEHA centre, which provided IT services to TE Services. The experiment register included information on the target population, treatment status, randomisation strata and entrepreneurs, as well as decisions and payments for firms that applied for the subsidy. The legislation detailed the access to and use of the data for conducting the experiment.

In the analysis phase, the evaluator's data access was secured by the legislation regarding the experiment. The tax authority provided outcome and background data to VATT, while the KEHA centre provided the experimental register, which was merged with the tax data. All data used in the analysis were pseudonymised. The evaluation used tax data from the Incomes Register, including monthly wage costs and employees, firm income tax returns, VAT returns and individual income tax returns. The extensive data allowed for an evaluation of the effects on outcomes at the firm, entrepreneur, and worker levels.

3.9 Evaluation

One important feature of the legislation was that it granted the evaluator access to all relevant register data held by public authorities to conduct the evaluation. It also outlined that the public authorities, including the tax authority, had to provide the data at no cost. This sped up the results.

Einiö and Nivala (2026) provide the results of the evaluation. According to their study, the subsidy had a positive and statistically significant effect on the proportion of firms that became employers. The proportion of firms that hire workers increased by 0.6 percentage points during the first six months (20% relative to the baseline) and 0.7 percentage points (11% from the baseline) over the whole subsidy period. They find effects of a similar magnitude on wage costs and the number of employees. Importantly, they find that the positive increase in the proportion of employers persists even after the end of the subsidy period, which suggests that the subsidy, while temporary, created permanent employers.

According to Einiö and Nivala (2026), the subsidy created 207 new employers and 138 permanent employers by the end of the observation period, making the direct subsidy cost of an employer and a permanent employer €18,200 and €27,300, respectively.^[29] In other words, every €1 in subsidies created €2.3 of labour input. In addition, according to their results, the average monthly number of employees increased by 0.006 during the subsidy period, corresponding to 207 individuals, which is equivalent to the number of new employers. The authors argue that because the subsidy had a permanent effect on hiring, it appears to have helped some non-employer firms overcome a recruitment threshold. However, while the subsidy had a statistically significant and, in relative terms, large effect on the incidence of hiring first employees and on employment, the overall employment effects are limited by the low take-up rate.

The study also finds that most of the hired employees were already employed, and therefore, the subsidy may have primarily increased the hiring of already employed people. If the subsidy merely reallocated individuals who would have been working regardless of the support, this may further reduce the effect on overall employment. Finally, the study does not show statistically significant effects on measures of firm growth other than employment, although the estimates for revenue and added value are positive, which is consistent with the positive employment effect.

29. A calculation of the *total* cost of a new employer should include administrative costs to the public administration and firms. Information on such costs is not available, but the expectation is that they will be fairly low relative to the total budget, due to the simple application process and the automatic payment arrangement, based on the Income Register, under which employers are obligated to file wage information regardless of the experiment.

3.10 Comparison of RCT results with previous work

In this section, we compare the recruitment subsidy experiment to the earlier regional first-employee subsidy programme. Table 1 summarises the treatments and key findings from these studies. The results differ greatly. The RCT had an economically and statistically significant effect on hiring probability, while the regional programme had no detectable effect. The take-up rate in the RCT is an order of magnitude higher than in the regional programme.

Both instruments covered wage costs. For the regional subsidy, the support rate was 30% in the first year of use, and 15% in the second. The recruitment subsidy support rate was 50% for one year. They are therefore comparable in size, but the recruitment subsidy is front-loaded compared to the earlier regional subsidy. The realised subsidy payments were of a similar magnitude for recipients. Hence, the support rate or financial benefits of the subsidy are unlikely to explain the differences in the results. In addition, in the RCT, the firms only had five months to hire their first employee to qualify for the treatment, compared to approximately four years in the regional programme. If anything, this should reduce the effect of the RCT compared to the regional programme.

The main difference between the results of the two studies is that the recruitment subsidy design led to a higher take-up rate among new employers, which in turn had a positive effect on becoming an employer. The main design differences are (i) flexibility of the employment contract in the RCT; (ii) lower administrative costs and timely subsidy payments in the RCT; and (iii) direct information letters sent to the treatment group in the RCT. Based on this evidence alone, it is impossible to disentangle which of these best explains the difference in the take-up rate. However, it is clear that when these design features are implemented together, as in the RCT, they can have a large impact on a programme's effectiveness. Nivala (2024) argues that restricting the regional first-employee subsidy to full-time contracts and potential information frictions reduced take-up. In addition, the regional programme had higher administrative costs. It is likely that a combination of these factors explains the difference in take-up rates.

Table 1. Comparison between the RCT and prior evaluation of the regional subsidy

	Recruitment Subsidy Experiment 2022	Regional First-Employee Subsidy 2007-2011
Study	Einiö and Nivala (2026)	Nivala (2024)
Research method	Randomised controlled trial	Regional difference-in-differences
Subsidy rule	50% total wage costs for 12 months, max €10,000	30% first-employee wage costs for the first year, 15% second year
Time to take up	5 months	3.5 to 4.5 years
Eligibility	No employees for 12 months, treatment group	No employees for 12 months, in eligible municipality, hires for permanent contract, at least 25 hours/week
Administration	Simple online application, automatic monthly payments	Simple application, apply for payments after realised wage costs twice a year
Information	Directly informed by letters	Spread through ELY Centres, firms may need to seek information
Estimated effect (std. err.)	0.006 (0.001)	-0.003 (0.005)
Effect relative to baseline	20%	-4%
Number of targeted firms	34,500	c. 40,000
Take-up rate in target population	1.5%	0.2%
Take-up rate among new employers	32%	1.9%
Average paid subsidy (EUR)	8,131	7,459

Note: Information and estimates gathered from referenced studies.

4 Recommendations

The parameter of interest to be estimated depends on the policy instrument and the research questions. Our general recommendation is to consider designing experiments intended to estimate the ITT effect, especially in cases of public policy interventions with incomplete take-up.

Planning is often needed before political decisions can be made about large-scale experiments. In such processes, it would be important to integrate, to a sufficient degree, research perspectives and analysis, e.g., when the policy design changes. In such situations, the experiment's power to detect effects may change, and it is therefore important to recalculate the statistical properties of the experiment.

A significant and demanding part of an experimental project's scientific work takes place during the design and pre-analysis phases, and therefore at this stage, it is advisable to avoid uncertainty regarding the continuation of the project. In the Recruitment Subsidy Experiment, the pre-planning, the scientific support provided during the preparation of the proposed legislation and the implementation of the treatment allocation, as well as the evaluation, were carried out in separate phases. This meant that the researchers experienced some uncertainty as to whether they would ultimately be able to evaluate the experiment. Our recommendation is that the planning and evaluation components of projects should be included as part of a single, comprehensive contract. This would allow the research side to be better integrated into the research programme in the long term. It would also create stronger incentives, as it ensures that the researchers can utilise the scientific outputs from their design and pre-analysis work.^[30]

One aspect not implemented in the Recruitment Subsidy Experiment was the collection of survey data to identify the factors that influence the low take-up rate for this type of support, and why some firms in the treatment group that hired employees chose not to use the subsidy. However, the project lacked sufficient resources to carry out such a survey. Our recommendation for future projects is that sufficient funding be allocated for surveys of this kind. It is important that such surveys take place after the end of the treatment period, including in projects where population registers are the primary data source.

30. When the comprehensive, single contract model is used, it may be necessary to grant the contracting authority the right to terminate the project so that subsequent phases are not carried out. This may be necessary if, for example, it emerges during the design phase that the project has a low level of feasibility.

5 Concluding remarks

The intended functioning of public policies is crucial for social and economic welfare. Randomised controlled trials are the gold standard for reliable assessments of the impacts of policy interventions. Yet governments often roll out new and expensive programmes with little evidence regarding their impacts, despite the fact that experimentation prior to full implementation provides significant opportunities (Muralidharan and Niehaus, 2017). The resulting lack of evidence on what works and what does not is therefore likely to reduce the effectiveness of aggregate public spending significantly.

The Recruitment Subsidy Experiment, as described in this article, studied how entrepreneurs respond to incentives for hiring their first employees. While there was a previous observational study on a similar policy instrument (Nivala, 2024), the current experiment provided new insights. Contrary to the previous results, Einiö and Nivala (2026) find that the subsidy programme had a significant effect on hiring rates among non-employer firms. Together, these two studies show how policy design matters for effectiveness. Duflo (2020) describes field experiments as a practice of policy: one experiment reveals whether a treatment worked; researchers and practitioners use that evidence to redesign the next treatment; and the new experiment again increases our understanding of what works. The Finnish Recruitment Subsidy Experiment is one example of such a practice, which builds on previous evaluations to develop and test an improved type of policy design.

We discussed issues that are likely to arise in the evaluation of many typical public policies with costly treatments and incomplete take-up. As a practical example of such an evaluation, we discussed the Recruitment Subsidy Experiment. We hope that this study will inspire similar large-scale experiments aimed at assessing and improving the effectiveness of a wide variety of public policies.

References

- Bhargava, S. and Manoli, D. (2015). Psychological frictions and the incomplete takeup of social benefits: Evidence from an IRS field experiment. *American Economic Review*, 105(11):3489–3529.
- Duflo, E. (2020). Field Experiments and the Practice of Policy. *American Economic Review*, 110(7):1952–1973.
- Duflo, E., Glennerster, R., and Kremer, M. (2007). Using randomization in development economics research: A toolkit. *Handbook of Development Economics*, 4:3895–3962.
- Einiö, E. and Nivala, A. (2026). Making employers: The effects of supporting first hires in a large-scale randomized experiment. SSRN Working Paper no. 6063771.
- Einiö, E., Nivala, A., and Nokso-Koivisto, O. (2021). Yksinyrittäjien rekrytointituen vaikutusten arviointi: Suunnitelma satunnaistetusta kontrolloidusta kokeesta. Työ- ja elinkeinoministeriön julkaisuja 2021:1.
- Finkelstein, A. and Notowidigdo, M. J. (2019). Takeup and targeting: Experimental evidence from SNAP. *The Quarterly Journal of Economics*, 134(3):1505–1556.
- FINLEX (2022). Laki rekrytointitukikokeilusta 20/2022. Helsinki, 14.1.2022.
- McConnell, B. and Vera-Hernandez, M. (2015). Going beyond simple sample size calculations: A practitioner's guide. IFS working paper w15/17, London.
- Muralidharan, K. and Niehaus, P. (2017). Experimentation at scale. *Journal of Economic Perspectives*, 31(4):103–124.
- Nivala, A. (2024). (No) Effects of Subsidizing the First Employee: Evidence of a Low Takeup Puzzle Among Firms. Working Papers 23, Finnish Centre of Excellence in Tax Systems Research.

Technical Appendix

Formally, the treatment group size for the second wave can be written as

$$r_2 = \hat{p}_1 + z_{\gamma}^* \frac{\hat{\sigma}_1}{\sqrt{n_1}} \quad (1)$$

where r_2 is the adjusted take-up rate used to calculate the number of offers in the second wave, $\hat{p}_1$ and $\hat{\sigma}_1$ are the estimates of the take-up rate and its standard error from the first wave, z_{γ}^* is the (one-sided test) critical value of the standard normal distribution corresponding to the chosen risk level γ for over-subscription, and n_1 is the number of observations in the first-wave treatment group. When over-subscription is to be avoided, this approach inflates the anticipated fraction of treatment-takers, lowering the number of treatment offers. For example, suppose that $\hat{p}_1 = 0.05$, $\hat{\sigma}_1 \approx 0.2179$, and $n_1 = 10,000$ (we note that for a binary variable, the standard error is a function of p in this setup: $\sigma = \sqrt{p(1-p)}$). If the researcher is willing to accept a risk level of over-subscription of 5%, we have $z_{0.05}^* \approx 1.6449$ and $r_2 \approx 0.05 + 1.6449 \cdot \frac{0.2179}{\sqrt{10000}} \approx 0.0536$. With a budget of B_2 left for the second wave and a cost per treatment of s , the second wave will offer the treatment to $n_2 = B_2 / (0.0536s)$ units.

One specific issue is how to determine the size of the first wave. The key threat arises from it being too large. However, in a large-scale setting, a relatively small first wave is typically sufficient to provide enough control to avoid over-subscription.^[31] Sometimes, previous studies or other existing measures can be helpful guides when setting the size of the first wave. However, this may be limited by the fact that even policies with similar goals may differ in design features that can have a considerable impact on take-up rates.

The size of the first wave can also be assessed with available register data on the incidence of taking the intended action, and by adjusting this information with an expected effect.^[32]

31. The first wave can also serve as a useful test of the experimental procedure, before scaling to the larger second wave.

32. The take-up rate is equal to the sum of the baseline probability of choosing the intended action (e.g. study, invest or employ) and the treatment effect. Formally, this is expressed as $p_1 = p_0 + \delta$, where p_0 is the baseline probability and δ is the effect of the treatment on the probability. The simplification here is to assume that all units that would take the intended action also take up the treatment. In reality, however, this is not necessarily the case.

Comments on Elias Einiö and Annika Nivala: Large-Scale Randomised Experiments with Population Data

Sofie Cairo

Einiö and Nivala (2025) summarise several important lessons from the Finnish Recruitment Subsidy Experiment (FRSE), an intervention they implemented in collaboration with the Ministry of Economic Affairs and Employment of Finland in 2022–23.

The objective was to turn companies without employees into employers by offering a temporary wage subsidy for their first employee, and the tool was a large-scale, nationwide randomised controlled trial (RCT). Based on evidence from previous Finnish wage subsidy policies, the risk of imperfect take-up was high.

The main challenge faced by the research team in designing the experiment was, therefore, how to deal with the uncertainty surrounding take-up vis-à-vis offering a very costly treatment, in other words, striking the right balance between undersubscription with the implied risk of an underpowered statistical analysis and oversubscription with the implied risk of a budget overrun.

The research team took several steps to overcome these challenges. First, they agreed with the agencies involved to implement a relatively flexible policy offering subsidies of up to 50% for hiring a part-time or full-time worker with either a temporary or permanent contract for up to 12 months, up to a ceiling of €10,000. However, the window for take-up was limited to four months.

Second, they took steps to maximise awareness of the policy among treated companies by sending invitation letters to participate in the RCT directly to the owners.

Third, they took steps to minimise the costs of compliance, defined broadly as any paperwork involved in signing up for the programme, obtaining refunds for wage costs from the programme, and ensuring that refunds were paid out promptly and automatically.

To overcome any remaining uncertainty about take-up rates, the authors settled on a staged rollout, in which the initial pilot phase cautiously invited a small proportion of the participating population to generate direct evidence of expected take-up before implementing the more costly main arm of the experiment.

The pilot stage proved crucial, as it turned out that more than half of the target population could be invited to participate in the programme with practically no risk of oversubscription due to very low take-up rates, even with direct invitations. As a result, during the main experiment, the researchers focused solely on achieving optimal group sizes for the control and treatment groups to ensure statistical power.

Based on the intervention, the authors derived insights into running population-wide RCTs in collaboration with official bodies in a Nordic setting, which I summarise and discuss below.

The authors propose that the Nordic countries offer an ideal setting for population-wide interventions that support smooth, efficient implementation. First, researchers and authorities can rely on a rich digital infrastructure comprising platforms for direct digital communication with, e.g., company owners, employees, job seekers, taxpayers, and other targeted population groups.

These online platforms facilitate the direct delivery of invitations to subsidy programmes and targeted information interventions. In some instances, platforms even require participants to have access to and regularly read emails to remain eligible for benefits (Cairo & Mahlstedt, 2022).

Relying on existing infrastructure has multiple advantages, including lower costs and shorter implementation and running times, as well as more efficient treatment uptake.

In the case of FRSE, a dedicated platform was created for fast, smooth sign-up, while the prompt, automatic payout of wage refunds to compliers was facilitated through an existing platform used by Finnish employers to report all employee wage costs.

While implementation costs, including invitation costs, were fairly low, the policy intervention clearly had high costs per treated complying unit, since it relied on subsidies of up to €10,000 per complying company. By contrast, the marginal cost for each invited unit, whether complying or not in the case of information treatments, approaches zero when existing platforms are used.

Second, access to administrative registers across the Nordic countries supports perfect stratified randomisation based on pre-intervention observables such as industry codes and company size. This guarantees balanced treatment and control groups.

Another aspect that could have been explored in the FRSE is general equilibrium effects. While the hiring subsidy generated only approximately 207 jobs across Finland,

related wage subsidies available to a larger target population could increase competition for qualified labour, resulting in higher overall wage costs. In addition, reallocating labour from more productive large companies to less productive sole proprietorships could lower overall economic growth.

The Nordic setting is ideal for exploring such spillover effects, thanks to the fine-grained geographical information in the registers, which allows for the delineation of local labour markets by location and industry. This implies that intervention designs can vary the degree of treatment across local labour markets and investigate the general equilibrium effects. One example of this is found in Altmann et al. (2022), which offered online job search advice to jobseekers and varied the proportion of those treated in local labour markets.

Access to administrative registers also facilitates the collection of evidence on outcomes at the evaluation stage of the intervention, such as wage costs, profits and revenues, and the number of employees over time. This is clearly an advantage over surveys, which are both costly and subject to attrition.

The authors note that they would have preferred to exclude hiring family members at the design stage. On the other hand, in a Nordic context with access to population registers, examining the proportion of hired individuals who are family members appears straightforward, as registers can usually be merged using personal identifiers. In fact, the registers facilitate a full analysis of employee characteristics, such as family ties, employment status at the time of hiring, wage income after hiring, compared to the year before. Similarly, the researchers can investigate whether employees hired with these subsidies are more likely to start their own businesses.

The FRSE intervention benefited from both policy support and institutional support. It was based on cooperation and collaboration with national agencies, and it had a special advantage: political preapproval for both the budget and design. It is impossible to exaggerate the importance of long-lasting alliances between research teams and public institutions for the successful implementation of field experiments.

Unfortunately, many public and private interventions depend on the goodwill and inquisitive nature of individuals within organisations, rather than on the organisations per se. This remains a source of vulnerability for interventions, especially given the research time horizon relative to the typical length of time public employees remain in post. On this aspect, the Nordic countries are not so different from other settings.

One limitation of the FRSE, as recognised by the authors, is that no presurvey or postsurvey was distributed to investigate perceived barriers and enablers to hiring amongst founders of non-employer companies. The distribution costs of such surveys are typically relatively low in the Nordic countries due to digital communication platforms, and the survey evidence would have supported a richer investigation of low

take-up mechanisms, as well as perceived benefits to the company, to innovation, and to employees from hiring subsidies.

Without such evidence, the authors are left to guess about the barriers to hiring in non-employer companies, such as uncertainty about worker productivity, lack of regulatory knowledge, or perhaps a preference for working alone. Evidence already exists for the positive effects of hiring via the offer of an infrastructure for posting jobs and screening candidates (Algan et al., 2020) and via the offer of information on employment regulations (Bertrand and Crépon, 2021).

The intervention also suffers from a classic policy-research dilemma. Policies proposed by politicians are often political compromises rather than economically efficient solutions. It is therefore commendable that official bodies in Finland approved testing the policy before its permanent adoption. Nonetheless, before designing the policy, it would have been advisable to clearly define the intended effects – on company founders, employees, and overall economic growth – and to assess them in light of existing evidence.

A central question is whether hiring subsidies are the best or cheapest policy to achieve the goal of turning non-employers into employers, and whether they may risk hindering economic growth. Essentially, size-based policies are likely to encourage the reallocation of labour from larger, more productive companies to smaller, less productive ones, thereby risking harm to overall economic activity. Also, limited liability companies are more likely to experience growth when hiring than non-employer sole proprietorships.

Regarding evaluation, the authors refer to their related paper, Einiö and Nivala (2026), which presents evidence that the policy increased hiring by 0.6 percentage points, a 20% increase relative to the baseline. This is equivalent to 207 additional employers during the subsidy period, and about 138 after the expiry of the subsidy, out of the 34,500 invited companies. As such, the policy was successful overall in achieving its goal of increasing hiring, although it indicates that the appetite for hiring among non-employers – even under a generous subsidy scheme – remains very low. Qualitatively, the effects on job creation are in line with prior evidence on general hiring subsidies, just as job-training programmes tend to increase employment to the benefit of jobseekers (Dahlberg et al., 2024; sectoral job training Katz et al., 2022).

The authors present a partial cost-benefit analysis showing that one euro of subsidy generated €2.3 worth of labour input. However, they also recognise that most employees were already hired before the treatment invitation and would presumably have continued in employment even without it.

Importantly, the authors' related study finds no evidence that companies without employees who benefit from the subsidy experience increased growth in profits, revenues, or added value. This is in line with the findings of Cockx and Desiere (2023) in a Dutch context.^[33]

Even when hiring subsidies target highly skilled labour such as scientific researchers, Kaiser and Kuhn (2016) find that while the number of employees increases temporarily while companies take part in the programme, there are no statistically significant effects on company performance, including on added value, net income, return on assets, wages per employee and labour productivity. Neither during the program nor in the post-program period.

Given the insignificant impact on company growth, it is worth asking who actually benefits from the FRSE policy? In the absence of pre- and post-survey evidence to shed light on the perceived benefits to the compliers, this question remains unanswered.

The FRSE study illustrates the benefits of the Nordic setting for the implementation of policy interventions in the form of population-wide field experiments. Particularly, the use of existing digital infrastructures and platforms can reduce or even eliminate the costs of rolling out population-wide interventions. The platforms can facilitate direct communication with individuals in the target population, support smooth implementation, and reduce sign-up and compliance costs.

Access to administrative registers enables stratified randomisation based on pre-intervention observables to ensure balanced treatment and control groups. The registers also facilitate the collection of evidence on outcomes at the analysis stage, such as company wage costs, profits and revenues, and the number of employees over time.

It remains the gold standard for politicians to work with research teams on RCTs to test the impact and effects of their policies before population-wide rollout, and the FRSE is a prime example of such a collaboration. However, while RCTs are an efficient tool for examining the bite and effects of policies, they may also expose weaknesses in the policies in question.

33. Existing evidence points to screening of workers as a prerequisite for positive company effects (Lombardi et al., 2018).

References

- Altmann, S., Glenny, A. M., Mahlstedt, R., & Sebald, A. (2022). *The direct and indirect effects of online job search advice*. IZA Working Paper.
- Bertrand, M., & Crépon, B. (2021). Teaching labor laws: Evidence from a randomized control trial in South Africa. *American Economic Journal: Applied Economics*, 13(1), 125–149.
- Cairo, S., & Mahlstedt, R. (2023). The disparate effects of information provision: A field experiment on the work incentives of social welfare. *Journal of Public Economics*, 226, 104987.
- Kaiser, U., & Kuhn, J. M. (2016). Worker-level and firm-level effects of a wage subsidy program for highly educated labor: Evidence from Denmark. *Research Policy*, 45(9), 1939–1943.
- Katz, L. F., Roth, J., Hendra, R., & Schaberg, K. (2022). Why do sectoral employment programs work? Lessons from WorkAdvance. *Journal of Labor Economics*, 40(S1), S249–S291.
- Le Barbanchon, T., Schmieder, J., & Weber, A. (2024). Job search, unemployment insurance, and active labor market policies. In *Handbook of labor economics* (Vol. 5, pp. 435–580). Elsevier.
- Lombardi, S., Nordström Skans, O., & Vikström, J. (2018). Targeted wage subsidies and firm performance. *Labour Economics*, 53, 33–45.

Comments on Elias Einiö and Annika Nivala: Large-Scale Randomised Experiments with Population Data

Antti Kauhanen

1 Introduction

This article offers a timely, practice-oriented guide to designing and executing national-scale randomised controlled trials (RCTs) in settings where population registers cover the relevant units of observation. Anchored in the Finnish Recruitment Subsidy Experiment (FRSE) conducted along with the Ministry of Economic Affairs and Employment, the study synthesises econometric and organisational lessons highly relevant to public policies characterised by costly treatments and imperfect take-up. It constitutes a valuable contribution to the growing literature and practice of experimentation at scale.

2 A Nordic advantage – if we remove bottlenecks

The Nordic countries' population registers provide a genuine comparative advantage for policy experimentation. The ability to measure outcomes for all target units in administrative data reduces the usual cost constraints on data collection, enabling large samples, high statistical power, and credible estimates of the policy's effect, which are crucial in contexts where participation is voluntary. Registers also facilitate planning (through ex ante power calculations using population data) and implementation (e.g., precise identification of the target population).

However, realising this advantage requires attention to three practical bottlenecks:

- Access to data: Streamlined and predictable access arrangements are necessary to conduct experiments on policy-relevant timelines.
- Timeliness of data: Outcome registers must be updated quickly enough to support interim analyses and timely reporting to policymakers.
- Cost of access: Excessive data access fees can deter experimentation and make learning slow.

3 What low take-up reveals – and how to learn more

The recruitment subsidy experiment highlights a central puzzle: take-up was only 1.5% in the target population, but 32% among companies that became new employers during the experiment. This pattern sheds light on where the main frictions arise and suggests that policy design and communication may be as important as financial incentives.

The article rightly emphasises that public policy experiments often estimate the impact of offering access to a treatment rather than forcing uptake. To translate offers into behaviour, the evaluation should budget for complementary process evidence to help explain non-acceptance by eligible companies that would plausibly benefit. Concrete steps include:

- Verifying whether letters were received and read (e.g., address accuracy checks and open-rate statistics for digital channels).
- Diagnosing perceived complexity or risk around eligibility, compliance, and audits.
- Solving timing and cashflow problems that might reduce benefits for small businesses.
- A/B testing message framing (e.g., simplification, deadlines, testimonials) to optimise applications without changing the underlying generosity.

These inexpensive additions can substantially improve the interpretation of experimental impact and guide rapid policy iterations under the same programme.

4 From quasi-experiments to RCTs: it is not just identification

One intriguing insight from the Finnish case is that the transition from observational evidence to randomised experiment was closely linked with a significant policy redesign. The revamped programme included clearer eligibility criteria, quicker, automatic payments, and direct outreach to eligible companies, elements that probably boosted both participation and impact. Consequently, the experiment revealed significant positive effects on hiring, contrasting with a previous observational study on a similar instrument that registered no effects. This highlights a crucial point: the differences between evaluations involve not only identification but also the policy framework. In Duflo's "practice of policy," experiments generate evidence that informs improved programme design, which is then tested in subsequent experiments. The Finnish case exemplifies this iterative approach.

5 Design–implementation–evaluation as a single contract

Another important lesson pertains to the institutional setup. The team collaborated with the ministry from the early stages of design through to implementation and evaluation. This included conducting power calculations using administrative data and providing feedback on the bill in Parliament and information letters before the legislation was enacted in January 2022. Researchers advocated for randomisation at the level of the entire target population, identified from the tax register, rather than from an applicant pool. This approach allowed for the estimation of population-wide impact, which is the most informative data for policymaking.

The “comprehensive contracting” model, which integrates design, implementation support, and evaluation, effectively aligns incentives, ensures that randomisation is feasible within legal and operational frameworks, and increases the likelihood that the results will be both policy-relevant and timely.

6 Beware general equilibrium effects

When experiments are extensive and national in scope, it is crucial to consider the possibility that the treatment may influence outcomes in the control group, potentially through labour-market displacement or wage pressure. In the context of recruitment subsidies, spillovers might be envisioned if treated companies increase wages or attract workers away from control companies. Although population registers enable the monitoring of market-level outcomes, credibly identifying spillovers is more complex and often necessitates specific design choices, such as cluster-level randomisation and varying saturation across regions, along with pre-analysis plans for measuring spillovers. This issue deserves explicit attention on the agenda for future national experiments.

7 Concluding remarks

The study compellingly demonstrates that conducting national-scale RCTs using population data is both feasible and highly informative for public policy, especially when treatments are expensive, and uptake is suboptimal. The FRSE experiment exemplifies how thoughtful design choices, such as clarity, speed, and proactive outreach, can enhance programme effectiveness and how experimentation can be seamlessly integrated into policymaking as a standard practice for improvement. To expedite learning, the Nordic policy ecosystem should intensify its focus on enabling infrastructure (data access, timeliness, and cost) and on institutional models that integrate design, implementation, and evaluation into a cohesive whole. This approach will produce more credible evidence, faster iterations, and improved public programmes.

About this publication

Nordic Economic Policy Review 2026: Advancing Policy Through Randomised Experiments

Antti Kauhanen, Roope Uusitalo, Simen Markussen, Marie Bjørneby, Kaisa Kotakorpi, Helena Keinänen, Jarno Tuimala, Olavi Tuomi, Ramin Izadi, Matti Sarvimäki, Morten Visby Krægpøth, Ingeborg Foldøy Solli, Johan Vikström, Jukka Mattila, Kari Hämäläinen, Stine Nyhus Larsen, Nikolaj Noer Poulsen, Michael Rosholm, Katrine Bønneland Tølbøll, Hannu Karhunen, Marjo Kyllönen, Kristiina Hannukainen, Elias Einiö, Annika Nivala, Sofie Cairo

Nord2026:001

ISBN 978-92-893-8459-9 (PDF)

ISBN 978-92-893-8460-5 (ONLINE)

ISBN 978-92-893-8432-2 (PRINT)

ISSN 1904-8092

<http://dx.doi.org/10.6027/nord2026-001>

© Nordic Council of Ministers 2026

Published: 6.5.2026

Cover photo: Ann-Sofi Rosenkvist/imagebank.sweden.se

Other photos: Advancing Evidence-Based Policy Photo: Sofia

Sabel/imagebank.sweden.se // The Norwegian Tax Experiment Photo: Unsplash.com //

Learning to Experiment Photo: Ann-Sofi Rosenkvist/imagebank.sweden.se //

Evaluating Active Labour Market Policies Photo: imagebank.sweden.se // Childhood and Adolescent Interventions in a Danish Setting Photo: Ann-Sofi

Rosenkvist/imagebank.sweden.se // Large-Scale Randomised Experiments with

Population Data Photo: Margareta Bloom Sandebäck/Imagebank.sweden.se

Layout: Louise Jeppesen

The Nordic Council of Ministers

The Nordic Council of Ministers is the official body for inter-governmental co-operation between Åland, Denmark, Finland, the Faroe Islands, Greenland, Iceland, Norway and Sweden.

In 2019, the Nordic prime ministers presented a vision of the Nordic Region as the most sustainable and integrated region in the world by 2030. The work of the Nordic Council of Ministers is designed to pursue that goal by making the Nordic Region green, competitive and socially sustainable.

Nordic Council of Ministers

Nordens Hus

Ved Stranden 18

DK-1061 Copenhagen

www.norden.org

Read more Nordic publications: www.norden.org/publications