NBER WORKING PAPER SERIES

USING AI TO GENERATE OPTION C SCALING IDEAS:
A CASE STUDY IN EARLY EDUCATION

Faith Fatchen
John A. List
Francesca Pagnotta

Working Paper 33924
http://www.nber.org/papers/w33924

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2025

We gratefully acknowledge the team effort making this research project possible. For excellent research assistance, we thank Nicholas Chen, Harikesan Balachandran, and Clare Smith. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed additional relationships of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w33924

Using AI to Generate Option C Scaling Ideas: A Case Study in Early Education
Faith Fatchen, John A. List, and Francesca Pagnotta
NBER Working Paper No. 33924
June 2025
JEL No. C9, C90, C91, C92, C93, C99

ABSTRACT

In recent years, field experiments have reshaped policy worldwide, but scaling ideas remains a thorny challenge. Perhaps the most important issue facing policymakers today is deciding which ideas to scale. One approach to attenuate this information problem is to augment traditional A/B experimental designs to address questions of scalability from the beginning. List 2024 denotes this approach as "Option C" thinking. Using early education as a case study, we show how AI can overcome a critical barrier in Option C thinking – generating viable options for scaling experimentation. By integrating AI-driven insights, this approach strengthens the link between controlled trials and large-scale implementation, ensuring the production of policy-based evidence for effective decision-making.

Faith Fatchen
University of Chicago
ffatchen@uchicago.edu

Francesca Pagnotta
University of Chicago
fpagnotta@uchicago.edu

John A. List
University of Chicago
Department of Economics
and Australian National University
and also NBER
jlist@uchicago.edu

1 Introduction

The experimental approach has emerged as a pivotal tool in modern economics, with roots in the lab (Smith 1962; Fiorina et al. 1978). Over the past few decades, experiments have been conducted in naturally-occurring settings providing policymakers with data that bridges theoretical insights and practical application (Harrison et al. 2004). The overarching thrust of this work has underscored the import of controlling the assignment mechanism to establish causality, offering a robust methodology to understand behavioral principles across different domains. This approach has been instrumental in addressing complex policy issues, from climate change to education, by generating empirical evidence with promises that policymakers can readily use (Burgess et al. 2021; Ferraro et al. 2013; Cotton et al. 2022).

But, have those promises been fulfilled? We would argue not often enough. One core reason is the scaling problem, or what has been denoted as the "Voltage Effect" (List 2022). The importance of scaling experimental findings cannot be overstated, as it addresses the critical challenge of generalizing small-scale findings to larger populations of people and situations. Across disciplines ranging from software development to medicine to education and beyond, between 50% to 90% of results will lose efficacy at scale (List 2022). For instance, an early childhood education program that shows promising results in a specific community may see its effects reduced to a fraction – potentially $1/10^{\text{th}}$ or $1/5^{\text{th}}$ – when implemented across an entire city, state, or country. This phenomenon arises for several reasons (List 2022), which all revolve around information problems. One core issue within this set of problems is determining which ideas should be scaled.

One recent approach to tackle this issue advocates flipping the traditional research model to an approach that, from the beginning, produces the type of policy-based evidence that the science of scaling demands (List 2024). This approach is denoted as "Option C" thinking, and augments efficacy trials by including relevant tests of scale in the original discovery process. Such an approach forces the scientist to naturally start with a recognition of the big picture: What information is necessary to provide scaling confidence? A key constraint naturally arises – determining the set of Option C ideas that should be considered.

In this study, we argue that AI can importantly complement Option C thinking. As our case study, we examine Option C thinking around the Chicago Heights Early Childhood Center (CHECC), which was established in 2010 in Chicago Heights, Illinois, as a preschool program designed to provide high-quality, tuition-free early education to children aged three to five (see Fryer et al. 2020; Castillo et al. 2024). While serving as a research hub to study educational interventions, it was also designed to explore the scalability of early education policies. The case study reveals that simple AI prompts can provide a wealth of Option C ideas that are both viable and efficient.

2 What is Option C Thinking?

As far back as Fisher 1935, it has been understood that the experimental backbone of social sciences is A/B testing, typically an efficacy trial pitting arm B (treatment) against arm A (control) to optimize intervention success in a controlled setting (List 2025). This "Petri dish" approach maximizes treatment effects but falters when guiding policy, as it tests ideas under ideal conditions rather than at scale (List 2024). To deliver robust policy advice, experiments must extend beyond this framework, incorporating real-world constraints to ensure interventions remain effective when broadly implemented, addressing the scalability challenge head-on.

According to List 2024, integrating efficacy tests with scalability assessments in the initial research design is both economically and scientifically efficient. Termed "Option C" thinking, this approach questions what additional data is needed beyond A/B testing to scale an idea effectively. List 2024 advocates including a scalable program version alongside the idealized A/B test, emphasizing that Option C thinking is more than adding a treatment arm – it is an experimental enhancement generating policy-relevant evidence for scaling from the outset. This shift moves beyond efficacy, embracing a broader perspective that considers real-world constraints, scaling factors, and the presence of mediation paths and moderators.

Figure 1: Adding Option C Thinking to A/B Testing

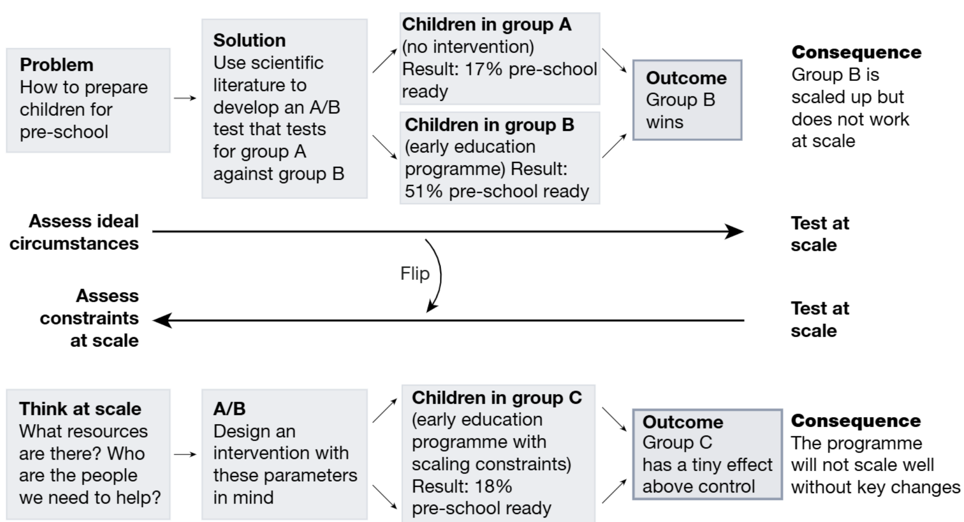

Figure 1, taken from List 2024, illustrates Option C thinking for an education program. In a standard A/B test, children are split into a control group (A, no intervention) and a treatment group (B, early education), with the main result showing that pre-school readiness rises from 17% (control) to 51% (treatment) under ideal conditions. Option C thinking, however, incorporates real-world scaling constraints, like using typical teachers. This scalable design, shown in the diagram's lower panel, yields only an 18% readiness rate – a mere 1%

increase over the control – revealing the program's limited cost-benefit viability at scale despite its efficacy in controlled settings. Accordingly, when the program is conducted with the constraints that it will face at scale, it does not seem to pass a cost–benefit test, even though it looked great in the efficacy test.

But, this thought experiment begs the question: How do I know which Option C ideas to try? Should I use economic theory? The previous literature? The five vital threats to scaling outlined in List 2024? Local knowledge? Pilot trials? Our answer to each of these is yes, and more.

3 Using AI to Generate Option C Thinking Ideas

A growing body of literature highlights AI's role in social sciences, particularly in experimental economics (Bail 2024; Chang et al. 2024; Charness et al. 2025). We explore AI's utility in generating Option C thinking ideas, which incorporate scalability into experimental design. For an early childhood education (ECE) program, a standard A/B test randomizes children into control and treatment groups. Option C thinking, however, requires addressing real-world scaling constraints. AI can assist by suggesting conditions, like typical teacher quality, serving as a low-cost tool to enhance idea generation for scalable experiments, despite its current limitations.

As our case study, we focus on the Chicago Heights Early Childhood Center (CHECC) project. As described in Castillo et al. 2024, CHECC was a large-scale field experiment in a low-performing urban school district in Chicago Heights, Illinois, where the 2020 median per capita income was approximately $17,546. Conducted from 2010 to 2014, CHECC recruited households with children aged three to five, originally comprising three groups: a preschool program, a parenting program, and a control group. For this analysis, we simplify to two groups: a treatment group receiving the preschool program and a control group, enabling a focused evaluation of the preschool intervention's scalability using Option C thinking.

Given AI's evolving capabilities, no standardized method exists for generating Option C thinking ideas. However, a structured approach can be followed: Train the AI model by uploading List 2024 to familiarize it with Option C thinking principles. Then, clearly define the experimental parameters. Finally, prompt the AI to generate Option C ideas that account for real-world scaling constraints. This process leverages AI as a flexible tool to propose scalable conditions, enhancing experimental design despite its current limitations.

As such, we propose the following three specific steps:

1. In the AI model of your choice, upload List 2024 to train the model on Option C thinking

2. For the program you want to generate Option C thinking ideas, clearly define:

(a) Research question(s)
(b) Setting
(c) Sample
(d) Treatment(s)
(e) Scaling objective
(f) Other relevant information

3. Given the above six items, ask the AI to generate Option C thinking ideas

Following this approach, for our CHECC case study, we define the following:

(a) Research question(s): What is the impact of preschool on skills in middle childhood and adolescence as measured by cognitive and non-cognitive skills?

(b) The setting: Low income school district in Chicago Heights, IL, USA

(c) The sample: Primarily low income families with children aged three to five

(d) Treatment(s): Children will be randomly assigned to receive the preschool program or the control group

(e) Scaling objective: We would like to scale this program across low-income populations in the U.S.

(f) Other relevant information: No other relevant information

3.1 Option C Results

This simple exercise revealed three reasonable Option C ideas that certainly merit further discussion by the research team. In terms of outcomes, AI produced the following Option C thinking ideas:

1. Idea 1: Typical Teacher Pool & Resources

   (a) Rationale: In the best-case (efficacy) trial, you might hire especially motivated, highly credentialed teachers or give them extra resources. At scale, it's far likelier you'll have to hire from a standard pool – potentially with less robust training, limited budgets, and varied levels of experience.

   (b) Option C Condition:

   - Teacher Recruitment & Training: Recruit and assign teachers under the normal district hiring process (i.e., limited salary caps, standard certifications).

- Resources & Budget Constraints: Provide typical classroom resources (e.g., standard teacher–student ratios, curriculum materials consistent with state mandates).

This first AI suggestion is consonant with the literature arguing for high quality teachers in early childhood education. Among others, Darling-Hammond 2000 and Andrew et al. 2024 address the import of teachers for students' learning and as a key means to augment human capital at young ages.

2. Idea 2: Large Class Size or Multi-Classroom Scaling

   (a) Rationale: At scale, classrooms often grow in size, or you must replicate many classrooms across multiple districts. Larger class sizes can reduce individualized attention and could impact results.

   (b) Option C Condition:

      - Expanded Enrollment: Instead of capping class sizes at, say, 15 children per teacher, have a larger ratio (like 20–25 per teacher), or operate multiple classrooms concurrently.
      - Standard Real-World Facilities: Use typical classroom space, with minimal renovations or specialized learning spaces.

   This recommendation is directly in line with the literature that explores how student teacher ratios affect value added. For example, the significance of class size has been underscored by a number of studies, including Hanushek 1999 and Hirsh-Pasek et al. 2004, which highlight its potential impact on programs outcomes.

3. Idea 3: Variation in Parental Engagement & Supports

   (a) Rationale: In the experiment, parents of low-income children may have self-selected into the program or been especially motivated. At scale, not all parents will engage equally (work schedules, transportation, competing life demands). This matters because parental involvement can strongly influence outcomes.

   (b) Option C Condition:

      - No or Minimal Parent Incentives: Compare outcomes if you do not provide any additional incentives or parent-focused resources.
      - Diverse Scheduling & Communication Approach: Offer the usual day-to-day communication (e.g., standard teacher notes, typical app or email updates) without extra home visits or specialized parent sessions.

   This final recommendation again follows the literature and represents a key consideration for scaling CHECC. The role of parents for children's development is explored in papers such as Carneiro et al. 2019, Cunha et al. 2008, Mahoney et al. 2007 and several others.

3.2 AI: The Beginning, Not Ending of Option C Ideation

The six-item list that we used – defining research question(s), setting, sample, treatment(s), scaling objective, and other relevant details – serves as the foundation for prompting AI to generate Option C thinking ideas, as outlined in List 2024. This structured approach ensures AI aligns with the experimental design's scalability goals. However, to enhance AI's creativity, researchers can enrich prompts with additional elements, such as specifying real-world constraints (e.g., budget limits or typical staffing) or requesting multiple scenario-based ideas. Examples of such prompts are provided in the Appendix (Section 5). Iteration is key when working with AI, as its outputs often require refinement to suit specific project needs. Researchers should experiment with prompt variations, adjusting specificity or context to elicit novel, actionable ideas.

Importantly, AI serves as a supportive tool, not a replacement for rigorous analysis. All AI-generated ideas must be critically evaluated, pairing them with a thorough literature review to ensure theoretical and empirical grounding. This involves assessing the feasibility of proposed conditions, like using standard teacher pools, against existing studies and scalability challenges identified in works like List 2024. Critical thinking is essential to filter out impractical suggestions and refine viable ones, ensuring alignment with the experiment's objectives.

By combining AI's ideation potential with disciplined scrutiny, researchers can effectively integrate Option C thinking into experimental designs, producing policy-relevant evidence that withstands the complexities of scaling. This iterative, critical approach is particularly vital given AI's current limitations, such as its tendency to generate generic or contextually misaligned ideas. For instance, in the Chicago Heights Early Childhood Center case, AI might suggest scaling conditions like larger class sizes, but researchers must verify these against local educational constraints and prior findings. By tailoring prompts and rigorously vetting outputs, researchers can leverage AI to enhance Option C thinking, bridging the gap between controlled trials and large-scale policy implementation.

4 Conclusion

A/B testing, the cornerstone of the experimental sciences, has provided key insights across every sector of modern economies. Within policy circles, the experimental approach is critical, yet recently research has shown many interventions experience a "voltage drop" when scaled, where the cost-benefit profile depreciates considerably, making it difficult to solve social problems effectively at larger scales. In some circles, this is denoted as the 'scale up' problem.

One solution to the scale up problem is to integrate scalability into initial experimental designs, denoted as "Option C" thinking in List 2024. Option C thinking is a methodological shift from traditional A/B testing, which focuses on efficacy under ideal conditions, to a design that incorporates scalability from the beginning. Beyond traditional efficacy tests, Option C thinking incorporates real-world constraints, such as typical teacher quality or larger class sizes, to assess programs at scale.

Using the Chicago Heights Early Childhood Center as an example, this paper explores the possibility of AI generating Option C ideas. We find that there is much to be optimistic about, as AI unduly enhances experimental economics by simulating scalable conditions. This approach ensures policy-relevant evidence from the start, bridging the gap between controlled trials and large-scale implementation.

References

Andrew, Alison et al. (2024). "Preschool Quality and Child Development". In: *Journal of Political Economy* 132.7, pp. 2304–2345. DOI: 10.1086/728744. URL: https://doi.org/10.1086/728744.

Bail, Christopher A (2024). "Can Generative AI improve social science?" In: *Proceedings of the National Academy of Sciences* 121.21, e2314021121. DOI: 10.1073/pnas.2314021121. URL: https://doi.org/10.1073/pnas.2314021121.

Burgess, Simon, Robert Metcalfe, and Sally Sadoff (2021). "Understanding the response to financial and non-financial incentives in education: Field experimental evidence using high-stakes assessments". In: *Economics of Education Review* 85, p. 102195. DOI: 10.1016/j.econedurev.2021.102195. URL: https://doi.org/10.1016/j.econedurev.2021.102195.

Carneiro, Pedro et al. (2019). *Parental Beliefs, Investments, and Child Development: Evidence from a Large-Scale Experiment*. IZA Discussion Paper 12506. Institute of Labor Economics (IZA). URL: https://docs.iza.org/dp12506.pdf.

Castillo, Marco et al. (2024). "Detecting drivers of behavior at an early age: Evidence from a longitudinal field experiment". In: *Journal of Political Economy* 132.12, pp. 3942–3977. DOI: 10.1086/731409. URL: https://doi.org/10.1086/731409.

Chang, Samuel et al. (2024). *12 Best Practices for Leveraging Generative AI in Experimental Research*. Working Paper 33025. National Bureau of Economic Research. DOI: 10.3386/w33025. URL: https://www.nber.org/papers/w33025.

Charness, Gary, Brian Jabarian, and John A List (2025). "The next generation of experimental research with LLMs". In: *Nature Human Behaviour* 9, pp. 833–835. DOI: 10.1038/s41562-025-02137-1. URL: https://doi.org/10.1038/s41562-025-02137-1.

Cotton, Christopher, Brent R Hickman, and Joseph Price (2022). "Affirmative Action and Human Capital Investment: Evidence from a Randomized Field Experiment". In: *Journal of Labor Economics* 40.1, pp. 157–185. DOI: 10.1086/713743. URL: https://doi.org/10.1086/713743.

Cunha, Flavio and James J Heckman (2008). "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation". In: *Journal of Human Resources* 43.4, pp. 738–782. DOI: 10.3368/jhr.43.4.738. URL: https://doi.org/10.3368/jhr.43.4.738.

Darling-Hammond, Linda (2000). *Solving the Dilemmas of Teacher Supply, Demand, and Standards: How We Can Ensure a Competent, Caring, and Qualified Teacher for Every Child*. National Commission on Teaching & America's Future.

Ferraro, Paul J and Michael K Price (2013). "Using Nonpecuniary Strategies to Influence Behavior: Evidence from a Large-Scale Field Experiment". In: *The Review of Economics* *and Statistics* 95.1, pp. 64–73. DOI: 10.1162/REST_a_00344. URL: https://doi.org/10.1162/REST_a_00344.

Fiorina, Morris P and Charles R Plott (1978). "Committee Decisions under Majority Rule: An Experimental Study". In: *American Political Science Review* 72.2, pp. 575–598. DOI: 10.2307/1954111. URL: https://doi.org/10.2307/1954111.

Fisher, Ronald A (1935). *The design of experiments*. 1st. Oliver & Boyd.

Fryer, Roland G Jr et al. (2020). *Introducing CogX: A New Preschool Education Program Combining Parent and Child Interventions*. Working Paper 27913. National Bureau of Economic Research. DOI: 10.3386/w27913. URL: https://www.nber.org/papers/w27913.

Hanushek, Eric A (1999). "Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigations of Class Size Effects". In: *Educational Evaluation and Policy Analysis* 21.2, pp. 143–163. DOI: 10.2307/1164297. URL: https://doi.org/10.2307/1164297.

Harrison, Glenn W and John A List (2004). "Field Experiments". In: *Journal of Economic Literature* 42.4, pp. 1009–1055. DOI: 10.1257/0022051043004577. URL: https://doi.org/10.1257/0022051043004577.

Hirsh-Pasek, Kathryn et al. (2004). "Does Class Size in First Grade Relate to Children's Academic and Social Performance or Observed Classroom Processes?" In: *Developmental Psychology* 40.5, pp. 651–664. DOI: 10.1037/0012-1649.40.5.651. URL: https://doi.org/10.1037/0012-1649.40.5.651.

List, John A (2022). *The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale*. Crown Currency.

— (2024). "Optimally generate policy-based evidence before scaling". In: *Nature* 626, pp. 491–499. DOI: 10.1038/s41586-023-06972-y. URL: https://doi.org/10.1038/s41586-023-06972-y.

— (2025). *Experimental Economics: Theory and Practice*. 1st. Forthcoming. University of Chicago Press.

Mahoney, Gerald and Bridgette Wiggers (2007). "The Role of Parents in Early Intervention: Implications for Social Work". In: *Children & Schools* 29.1, pp. 7–15. DOI: 10.1093/cs/29.1.7. URL: https://doi.org/10.1093/cs/29.1.7.

Smith, Vernon L (1962). "An Experimental Study of Competitive Market Behavior". In: *Journal of Political Economy* 70.2, pp. 111–137. DOI: 10.1086/258609. URL: https://doi.org/10.1086/258609.

5 Appendix

Below are three different prompts that can be used to generate Option C thinking ideas with AI. Prior to providing a prompt, upload List 2024 to train the AI model of choice on what Option C thinking is. These prompts should be viewed as guidelines, rather than rules, on how to ask AI for Option C thinking ideas. As such, each prompt should be adjusted to better suit one's experimental design and aims.

5.1 Prompt 1

I need you to generate [number] Option C thinking ideas, which is the concept outlined in the paper I previously uploaded, for a research project I plan to conduct. To generate Option C thinking ideas you will need to think about the constraints the idea will face at scale, the key factors that can impact whether or not the idea scales, as well as the mediators and moderators that could be at play in the experimental setting outlined below. For each Option C thinking idea, please provide:

1. A detailed description of the proposed idea
2. A detailed description on how to implement the proposed feature into the experimental design
3. A detailed description of the specific mechanism that the proposed idea is getting at
4. A detailed description of how the proposed idea differs from my treatment arm(s)
5. A detailed description of why including this feature in the experimental design will help inform whether the intervention will work at scale

In terms of my experimental design and setting:

1. The research question is ...
2. The experiment will be conducted in ...
    - Provide details on the setting (e.g., location of schools where the experiment will take place)
3. The participants of the experiment will be ...
    - Provide details on the participants, such as their age, gender, income-level, etc.,
4. The control and treatment arms are ...
    - Provide details on the most salient features of each treatment arm

5. My scaling objective is ...
    - Provide details on the how the intervention being tested will be scaled (e.g., the aim is to scale the education program to all schools in district X)
6. Other relevant considerations for the experiment are ...
    - Provide details on any experimental design feature you deem salient. This could include the randomization mechanism used, possible moderators and mediators, etc.,

5.2 Prompt 2

Identical to Prompt 1, but in addition to providing the six experimental design details outlined above, AI is given a full pre-analysis plan (PAP). AI is informed about the nature of the uploaded document and its content. Once again, since these prompts should be viewed as guidelines, one should experiment with uploading PAPs of vary depth to see what works best for one's context.

5.3 Prompt 3

Identical to Prompt 1, but greater emphasis is placed on using creativity to generate Option C thinking ideas. The following can be added at the end of Prompt 1: When generating Option C thinking ideas think creatively about:

- Isolating components of the intervention that are key for scaling and Option C thinking
- Testing contextual factors that are relevant for scaling and Option C thinking
- Varying the delivery method(s) and/or intensity of the intervention to account for constraints that might be faced at scale
- Examining interaction effects that are relevant for scaling and Option C thinking
- Challenging assumptions in our current design

Be bold in your suggestions – I'm looking for ideas that will reveal surprising insights about the constrains that I must account for when scaling my idea.