AI and Productivity in Canada

Hans Martinez Martin Petrin

2025-04-28

Main Objective

The goal of the experiment is to measure whether using a generative AI tool improves worker productivity. The aim is to compare the outputs of employees who have access to the tool against those who do not.

To obtain reliable causal estimates, employees with varying levels of skill will be randomly assigned to either an AI-access group or a no-access group. Workers will complete tasks of different levels of difficulty, allowing us to study both the overall effect of AI access and how effects vary across tasks and skill levels (Brynjolfsson et al. 2023; Gambacorta et al. 2024; Dell'Acqua et al. 2023).

Productivity will be measured by metrics such as the volume of output, task completion time, and independent quality assessments (e.g., supervisor ratings or standardized rubrics).

We will calculate the sample size necessary to detect meaningful productivity differences, ensuring sufficient statistical power. Based on preliminary assumptions, we estimate needing approximately 72 employee-weeks of data to achieve 80% power at a 5% significance level. Depending on the number of available employees, the experiment may need to run between one and four weeks to collect sufficient observations.

Participation in the experiment will be voluntary, and employees will be informed that their performance data will be used only for research purposes, without any negative consequences. Results of the study would be anonymized and sensitive information from the firm will remain confidential.

Secondary Objectives

Secondary objectives might be included in the projects depending on the available resources and the final scope of the experiment.

A second objective might be to evaluate different ways employees might use generative AI. We distinguish two main interaction approaches (Handa et al. 2025):

- Direct use: Employees use AI-generated outputs quickly and with minimal modifications.
- Collaborative use: Employees iteratively interact with the AI, refining outputs and incorporating their own insights.

For example, a direct user might ask the AI to draft a full report and submit it with few edits, while a collaborative user would use generative AI to obtain feedback on a report and improve it over several rounds of suggestions and edits.

A third objective could be to study the potential trade-off between productivity gains and employee skill development. For example, considering classifying tasks according to the Bloom's Taxonomy of cognitive processes (LW et al. 2001). If employees delegate higher-order cognitive tasks —as classified — (e.g., creating, analyzing) to AI, although they might become more productive in the short term, workers hinder their development of independent problem-solving skills. Conversely, delegating only lower-order tasks (e.g., remembering, understanding) might allow skill development to continue, but the gains in productivity might not be as large as delegating higher order cognitive tasks. Why? Understanding before creating a solution might take more time, than simply asking for the solution and testing if it works. On the other hand, debugging a solution that the worker does not understand might also take considerable time.

In any case, this potential trade-off could be critical for firms because long-term growth and innovation might critically depend on maintaining a workforce with strong critical thinking and problem-solving abilities.

To address these secondary objectives, the experiment will randomly assign employees not only to AI-access and no-access groups but also to different combinations of:

- Interaction approaches (direct vs. collaborative)
- Cognitive task levels (higher-order vs. lower-order)

Resources Needed

Ideally, the experiment would take place in an organization that has not yet introduced generative AI but plans to do so.

The experiment would focus on a group of employees performing similar tasks with measurable outputs. For example:

- Written reports
- Coding assignments
- Customer service responses
- Data entry or categorization

Employees may differ in expertise or skill level. Baseline data collection would occur before the intervention to measure:

- Existing skill levels (e.g., experience, performance history)
- Historical output for comparable tasks

The experimental phase would last four to six weeks, depending on operational constraints and sample size needs.

Next Steps

1. Define the project scope and finalize objectives.
2. Select a group of employees and identify the tasks.
3. Randomly assign employees to control (no AI) and treatment (AI access) groups, along with interaction approach and cognitive task level variations.
4. Measure and record outputs during the experimental period.
5. Analyze results to evaluate the impact of AI use, interaction style, and task delegation level.
6. Share AI access with all employees after the experiment concludes.

If the enterprise is already using a generative AI tool, the experiment could be reversed: temporarily restrict AI access for a group of employees to measure the effect of *removing* AI assistance on productivity.

References

Brynjolfsson, Erik, Danielle Li, Lindsey R Raymond, Daron Acemoglu, David Autor, Amittai Axelrod, Eleanor Dillon, et al. 2023. "Generative AI at Work," April. https://doi.org/10.3386/W31161.

Dell'Acqua, Fabrizio, Edward McFowland III, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani. 2023. "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." *SSRN Electronic Journal*, September. https://doi.org/10.2139/SSRN.4573321.

Gambacorta, Leonardo, Han Qiu, Shuo Shan, and Daniel M Rees. 2024. "Generative AI and Labour Productivity: A Field Experiment on Coding." www.deor.org.cn.

Handa, Kunal, Drew Bent, Alex Tamkin, Miles McCain, Esin Durmus, Michael Stern, Mike Schiraldi, et al. 2025. "Anthropic Education Report: How University Students Use Claude." April 8, 2025. https://www.anthropic.com/news/anthropic-education-report-how-university-students-use-claude.

LW, Anderson, Krathwohl DR, Airasian PW, Cruikshank KA, Richard Mayer, Pintrich PR, J. Raths, and Wittrock MC. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives.*

Appendix: Example of Power Calculation

Assumptions:

- Outcome: Number of tasks completed per week.
- Typical output without AI: 10 tasks/week.
- Expected output with AI: 12 tasks/week (+20% gain).
- Standard deviation (SD): 3 tasks.
- Significance level: 5% ($\alpha = 0.05$).
- Desired power: 80% ($\beta = 0.2$).

Sample Size Formula:

To compare two independent means:

$$n = \left(\frac{2 \times (z_{1-\alpha/2} + z_{1-\beta})^2 \times \sigma^2}{\Delta^2} \right)$$

Where:

- $z_{1-\alpha/2} = 1.96$ (for 95% confidence)
- $z_{1-\beta} = 0.84$ (for 80% power)
- $\sigma = 3$ (standard deviation)
- $\Delta = 2$ (difference in means: 12 - 10)

Substituting values:

$$n = \left(\frac{2 \times (1.96 + 0.84)^2 \times 9}{4} \right)$$

$$n = \left(\frac{2 \times (2.8)^2 \times 9}{4} \right)$$

$$n = \left(\frac{2 \times 7.84 \times 9}{4} \right)$$

$$n = \left(\frac{141.12}{4} \right) = 35.28$$

Thus, approximately 36 employees per group are needed.

Experiment Duration Options

Situation	Number of Employees	Experiment Duration Needed
Ideal	72 employees	1 week
Fewer employees	36 employees	2 weeks
Very few employees	18 employees	4 weeks

Thus, depending on availability, the experiment duration may vary to reach the necessary number of employee-weeks.

Employee Assignment Flow