# Psychological Insights into Decisions Relevant to Public Policy

Thesis by
Marcos Felipe Nazareth Gallo

In Partial Fulfillment of the Requirements for the
Degree of
Social and Decision Neuroscience


CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2025
Defended August 21, 2024

Marcos Felipe Nazareth Gallo
ORCID: 0000-0002-8227-2661

# ACKNOWLEDGEMENTS

Doing a PhD has been an intricate dance of self-discovery and collaboration, a delicate balance between solitude and connection. It has been like a choreography between me and my own shadow (see Houston & Hastings, 2022). Like the fall and recovery of Humphrey and Limón (Craine & Mackrell, 2010a, 2010b, 2010c), it kept my ADHD on its toes, constantly challenging me to find new rhythms and return to my breath.

That shadow was the one who propelled me out of Brazil, with the corner of my eyes still capturing that which I left behind with only the intent to help it. That shadow led me to daily Mandarin classes, moved my feet as I stepped onto Thai soil, and kept me from walking out of my first class in 高级微观经济学(advanced microeconomic theory) even though I couldn't understand a word of that strong northern accent. The same shadow prophetically wrote "Caltech Neuroscience 2018" on a blackboard kissed by a morning Shanghainese sun.

But the PhD cast a light on that shadow, and "by lights do shadows grow" (Houston & Hastings, 2022). I didn't know I had depression until I started writing a thesis. And what held me together? A loving partner, Sarah Quan, now my wife and mother of my daughter, Lucia—two women who give me joy and reason to live: 我爱你们. To my loving family and friends, who may not have understood what I was going through but believed in me: Obrigado Mãe, Pai, Ana, Vanda, Paulo e Thaís. Thank you to the Pages, the Quans, Janee, Josh, Moonsung, Dev, and Michaela. And to a fantastic therapist who methodically guided me as I performed repeated emotional relevés, who led me to self-leadership in my complex Internal Family Systems: thank you, Cheryl Youngblood.

Then, there are all those who granted me the time and space to find my footing when I stumbled. Thanks to Colin Camerer and Caltech administrators for their patience and care during these challenging times. Thanks to my thesis committee—John O'Doherty, Colin Camerer, Dean Mobbs, and Sera Linardi—for their invaluable guidance and support. I also sincerely appreciate the funding bodies that made this work possible: the Chen Institute, NSF, and the Haynes Foundation.

Along the way, others taught me their best steps, mentoring me in the finest scholarship and life lessons. Thank you to Carina Hausladen, Ming Hsu, Anna Jenkins, and Angela Duckworth. Thank you to those who taught me much by allowing me to

mentor them: Ben Manning, Kavya Rajagopalan, Sean Hu, Beatrice Maule, Mario Paiva, and Alec Guthrie. Thank you to Elizabeth Wallace for reading these pages and offering feedback, and Harshita for keeping me on track. And thank you to those who, as well as their guidance, provided a sense of community: Ralph Adolphs, Antonio Rangel, and my fellow SDN students.

The words of Brazil's greatest poet ring ever true (Andrade, 2015, pp. 40-41):

> Por isso sou triste, orgulhoso: de ferro.
> Noventa por cento de ferro nas calçadas.
> Oitenta por cento de ferro nas almas.
> E esse alheamento do que na vida é porosidade e comunicação.
>
> That's why I'm sad, proud, iron to the core.
> Ninety percent iron sidewalks.
> Eighty percent iron souls.
> And estrangement from all that's porous and communicates in life.

So I thank the Master Choreographer, the God who gave me the iron of determination and resilience. When I became estranged during the most challenging times, He placed my feet on solid ground to try again.

Perhaps my shadow, or "the longing to love, which cripples my work" (Andrade, 2015, p. 41), will always be with me wherever I go. But I hope I've learned to dance. Thank you all for teaching me.

**References**

Andrade, C. D. d. (2015). *Multitudinous heart: Selected poems : A bilingual edition* (R. Zenith, Trans.; First edition). Farrar, Straus; Giroux.

Craine, D., & Mackrell, J. (2010a). Fall-recovery. In *The oxford dictionary of dance* (2nd ed, p. 161). Oxford university press.

Craine, D., & Mackrell, J. (2010b). Humphrey, doris. In *The oxford dictionary of dance* (2nd ed, p. 223). Oxford university press.

Craine, D., & Mackrell, J. (2010c). Limón, josé. In *The oxford dictionary of dance* (2nd ed, p. 276). Oxford university press.

Houston, J., & Hastings, B. W. (2022, November 4). Dancing with my shadow [song]. In *Benjamin william hastings*. Capitol CMG/Invorto.

# ABSTRACT

This dissertation explores the integration of cognitive and behavioral sciences insights into policy-relevant domains, focusing on labor market discrimination, online teaching habits, and digital math education. The research comprises three studies:

1. An application of reinforcement learning models to analyze teachers' decision-making processes on the Zearn online math-teaching platform. This study demonstrates how computational models derived from computational psychology can capture complex, adaptive teaching behaviors and their impact on student outcomes.

2. A two-phase study combining data exploration with a large-scale field experiment to design and evaluate behavioral interventions for improving student learning outcomes on the Zearn platform. This research showcases the potential of data-driven approaches in developing effective educational interventions.

3. A meta-analysis of experimental correspondence studies investigating discrimination in North American labor markets. This study examines how perceptions of warmth and competence impact callback rates, providing insights into the mechanisms underlying discrimination.

This work demonstrates the feasibility and value of bridging cognitive and behavioral sciences with policy-making through innovative models and methodologies. The findings presented in this dissertation contribute to a more comprehensive understanding of how these fields can collaboratively tackle complex challenges in discrimination, education, and digital learning environments. Additionally, this research establishes a groundwork for future studies at the intersection of cognition, behavior, and public policy.

# PUBLISHED CONTENT AND CONTRIBUTIONS

Duckworth, A. L., Ko, A., Milkman, K. L., Kay, J. S., Dimant, E., Gromet, D. M., Halpern, A., Jung, Y., Paxson, M. K., Zumaran, R. A. S., Berman, R., Brody, I., Camerer, C. F., Canning, E. A., Dai, H., Gallo, M., Hershfield, H. E., Hilchey, M. D., Kalil, A., . . . Bulte, C. V. D. (2025). A national megastudy shows that email nudges to elementary school teachers boost student math achievement, particularly when personalized. *Proceedings of the National Academy of Sciences*, *122*(13). https://doi.org/10.1073/pnas.2418616122
M.G. designed three treatments arms.

Gallo, M., Hausladen, C. I., Hsu, M., Jenkins, A. C., Ona, V., & Camerer, C. F. (2024). Perceived warmth and competence predict callback rates in meta-analyzed North American labor market experiments (T. Otterbring, Ed.). *PLOS ONE*, *19*(7), e0304723. https://doi.org/10.1371/journal.pone.0304723
M.G. participated in the conception of the project, designed and performed the research, analyzed the data, and participated in the writing of the manuscript.

Buyalskaya, A., Gallo, M., & Camerer, C. F. (2021). The golden age of social science. *Proceedings of the National Academy of Sciences*, *118*(5), e2002923118. https://doi.org/10.1073/pnas.2002923118
M.G. performed research, analyzed data, and participated in the writing of the manuscript.

Lin, C., Keles, U., Tyszka, J. M., Gallo, M., Paul, L., & Adolphs, R. (2020). No strong evidence that social network index is associated with gray matter volume from a data-driven investigation. *Cortex*, *125*, 307–317. https://doi.org/10.1016/j.cortex.2020.01.021
M.G. helped with assembling data, preregistration, and carrying out a literature review for the introduction and discussion sections of the paper.

# TABLE OF CONTENTS

# LIST OF ILLUSTRATIONS

# LIST OF TABLES

# NOMENCLATURE

**Boosts.** Hints or further explanations provided by Zearn when students struggle with a problem.

**Callbacks.** Any response from an employer expressing interest in a particular candidate..

**Competence.** How capable a person is of acting on their intentions.

**ICA.** Independent Component Analysis.

**Intersectionality.** The interconnected nature of social categorizations such as race, class, and gender, creating overlapping systems of discrimination or disadvantage.

**Learning rate.** A parameter that determines how much new information overrides old information in reinforcement learning models.

**MDP.** Markov Decision Process.

**Megastudy.** A large-scale study involving multiple interventions tested simultaneously.

**NMF.** Non-negative Matrix Factorization.

**PCS.** Predicting Context Sensitivity.

**Q-value.** The expected cumulative future reward for taking an action in a given state.

**SCM.** Stereotype Content Model.

**Tower Alerts.** Notifications sent to teachers when a student struggles with a specific concept on the Zearn platform.

**Tower of Power.** An assessment feature in Zearn that presents students with challenging problems at the end of each lesson.

**Warmth.** The perception of how good or bad another person's intentions are.

**Zearn.** A digital platform for mathematics education used by approximately 25% of elementary school students in the United States.

*Chapter 1*

# INTRODUCTION

Governments worldwide grapple with discrimination and educational challenges. This dissertation presents research that offers examples of harnessing the power of cognitive and behavioral sciences to create data-driven solutions and positively influence behavior. The studies included in this thesis explore how insights from these disciplines can be applied to policy-relevant domains, focusing on labor market discrimination, online teaching habits, and digital math education.

In this introduction, I provide a brief background to the studies, explaining the context and importance of the research. Next, I outline the problem statements, followed by the research questions that guide this investigation. I then provide an overview of the methodology used across the three studies. Finally, I discuss the expected results and set out the structure of the subsequent chapters.

## 1.1 Background to the Studies

Social science research in the 21st century is experiencing an unprecedented transformation, marking its "Golden Age" (Buyalskaya et al., 2021). This era has brought forth interdisciplinary groups that break traditional academic boundaries, recognizing the importance of diverse perspectives in tackling complex social issues. Such dynamics propel social science toward tackling some of our time's most intricate and pressing challenges (Buyalskaya et al., 2021). The growth in interdisciplinary research within social science is also demonstrated by the increase in multi-investigator grants funded by agencies such as the NSF (National Academy of Sciences Staff et al., 2005), which underscores a systematic shift towards valuing collaborative approaches to address the increasing complexity of social phenomena.

The growing emphasis on interdisciplinary research in the social sciences provides a fertile ground for integrating cognitive and behavioral science into public policy. This integration has seen significant developments in recent years. The cognitive revolution in psychology, emphasizing mentalistic explanations for behavior (Miller, 2003), marked a paradigm shift in understanding human decision-making. This was followed by the emergence of behavioral economics, which challenged traditional economic models by incorporating psychological insights into economic decision-

making (see Buyalskaya et al., 2021). More recently, the concept of nudges has gained traction in policy circles, demonstrating how subtle changes in choice architecture can influence behavior without restricting freedom of choice (Thaler & Sunstein, 2003, 2009). The establishment of nudge units in governments, such as the UK's Behavioural Insights Team in 2010 further exemplifies the growing influence of behavioral science in policy-making (Halpern, 2015). These developments have paved the way for more evidence-based policy interventions that take into account the complexities of human behavior and decision-making processes.

The field of education has seen rapid shifts towards online and digital learning platforms, a trend accelerated by the COVID-19 pandemic (Di Pietro, 2023; Meeter, 2021). This transition has presented both opportunities and challenges for teachers and students alike (Alabdulaziz, 2021; Morrison et al., 2019). Adaptive online learning platforms have emerged as a promising strategy to mitigate adverse effects on learning, particularly in mathematics (Meeter, 2021; Ran et al., 2020). The Organisation for Economic Co-operation and Development's 2017 Report on Behavioural Insights and Public Policy highlighted how behavioral interventions, such as text message reminders, have been used to improve educational outcomes (OECD, 2017, p. 95-104).

Furthermore, platforms like Zearn have emerged as powerful tools for personalized learning. These platforms generate vast amounts of data on teacher and student behaviors, offering unprecedented opportunities for understanding and optimizing the learning process (Hershcovits et al., 2020; Salazar et al., 2007). However, effectively leveraging this data to improve educational outcomes remains an ongoing challenge (Al-Shabandar et al., 2018; Qiu et al., 2022; Shin & Shim, 2020).

In the realm of labor market discrimination, despite legal protections, subtle biases continue to influence hiring decisions (Bertrand & Duflo, 2017). Correspondence studies have been a primary tool for investigating this phenomenon, revealing persistent disparities in callback rates based on factors such as race, gender, and age (Lippens et al., 2023; Quillian & Lee, 2023). These studies have documented common patterns of discrimination across different social categories, with race often emerging as the strongest factor (Lippens et al., 2023). The Stereotype Content Model, proposed by Fiske et al. (2007), offers a framework for understanding these biases through the dimensions of warmth and competence.

The potential of interdisciplinary approaches in addressing these challenges is substantial. By combining insights from various fields, researchers can develop more

comprehensive solutions that account for the underlying cognitive and social factors at play (Buyalskaya et al., 2021; Pernu & Elzein, 2020). This approach allows for a more nuanced understanding of complex social phenomena and can lead to more effective policy interventions (Haushofer & Fehr, 2014; Stiglitz et al., 2019).

### 1.2 Statement of the Problem

Despite the growing integration of behavioral science into public policy, significant challenges remain in effectively applying these insights to complex societal issues. Specifically, this dissertation raises the key problem areas: a) In online education, vast amounts of data are available, but effective methodologies for translating these data into actionable insights and effective interventions are still lacking. Two of the studies included here aim to develop such methods; b) As previously noted, despite being well-documented, the underlying mechanisms driving labor market discrimination are not yet fully understood (Lippens et al., 2023). One of the studies in this dissertation aims to deepen our understanding of these mechanisms.

### 1.3 Research Questions

These studies aimed to address the following overarching question: In what ways can insights from behavioral and cognitive sciences be effectively applied in policy-relevant domains? This broad question was explored through three specific studies, each addressing a distinct aspect of the problem:

1. To what extent can reinforcement learning models capture and predict complex teaching behaviors in online math education platforms, and in what ways can these insights inform the design of digital learning environments and teacher support systems?

2. In what ways can unsupervised machine learning techniques be combined with field experimentation to develop and evaluate effective behavioral interventions for improving student learning outcomes in digital math education?

3. In what ways do perceptions of warmth and competence influence callback rates in labor market discrimination, and can this understanding inform more effective anti-discrimination policies?

These research questions were designed to address gaps in current studies and provide actionable insights for policy-makers and practitioners in the fields of labor market regulation, teacher education, and educational technology design.

## 1.4 Methodology Overview

The methodology employed in this dissertation flowed from an interdisciplinary approach that integrates methods and insights from behavioral science and public policy. The research combined various techniques to provide a comprehensive understanding of the complex phenomena under study. More precisely, each of the three distinct studies employed a unique methodological framework:

Study 1 applied reinforcement learning models, adapted from cognitive neuroscience and computational psychology, to analyze teachers' decision-making processes on the Zearn online math-teaching platform. This innovative approach combined computational modeling with analysis of large-scale behavioral data to capture the complex, adaptive nature of teaching strategies in digital environments.

Study 2 utilized a two-phase approach, combining unsupervised machine learning techniques with a large-scale field experiment. In the first phase, Independent Component Analysis was used to identify key dimensions of teacher behavior associated with student success. These insights then informed the design of behavioral interventions, which were evaluated through a randomized controlled trial involving over 140,000 teachers.

Study 3 employed a meta-analysis of experimental correspondence studies to investigate the role of perceived warmth and competence in labor market discrimination. This quantitative approach allowed for the synthesis of findings across multiple studies, providing a robust understanding of the mechanisms underlying hiring discrimination.

Data collection methods included surveys, meta-analysis of published studies, and extraction of user activity data from the Zearn platform. Analysis techniques ranged from traditional statistical methods to machine learning algorithms and computational modeling.

## 1.5 Hypotheses

Based on the existing literature and preliminary analyses, the following results were anticipated:

For Study 1, it was hypothesized that reinforcement learning models would effectively capture and predict complex teaching behaviors on the Zearn platform. These models were expected to outperform traditional methods in explaining teachers' adaptive strategies in digital learning environments.

In Study 2, it was anticipated that data-driven behavioral interventions, designed based on insights from unsupervised machine learning, would lead to significant improvements in student learning outcomes. Specifically, interventions focusing on teacher empathy and strategic engagement patterns were expected to increase student lesson completion rates.

For Study 3, it was expected that perceptions of warmth and competence will significantly predict callback rates in labor market experiments. Specifically, applicants perceived as high in both warmth and competence were expected to receive more callbacks, while those perceived as low in either dimension were assumed to be likely to face greater discrimination.

These expected results, if confirmed, were anticipated to provide strong support for the value of integrating behavioral science insights into policy-relevant domains.

### 1.6 Structure of the Thesis

The remainder of this dissertation is organized as follows:

Chapter 2 details the application of reinforcement learning models to analyze teachers' decision-making processes on the Zearn platform.

Chapter 3 describes the two-phase study combining unsupervised machine learning and field experimentation to develop and evaluate behavioral interventions for improving student learning outcomes.

Chapter 4 presents the meta-analysis of correspondence studies, examining the role of perceived warmth and competence in labor market discrimination.

Chapter 5 synthesizes the findings from all three studies, discusses their implications for policy and practice, and outlines directions for future research in the integration of cognitive and behavioral science into public policy.

## References

Alabdulaziz, M. S. (2021). Covid-19 and the use of digital technology in mathematics education [31 citations (Crossref) [2024-01-03]]. *Education and Information Technologies*, *26*(6), 7609–7633. https://doi.org/10.1007/s10639-021-10602-3

Al-Shabandar, R., Hussain, A. J., Liatsis, P., & Keight, R. (2018). Analyzing learners behavior in MOOCs: An examination of performance and motivation using a data-driven approach. *IEEE Access*, *6*, 73669–73685. https://doi.org/10.1109/access.2018.2876755

Bertrand, M., & Duflo, E. (2017). Field experiments on discrimination. *Handbook of Field Experiments*, *1*, 309–393. https://doi.org/10.1016/bs.hefe.2016.08.004

Buyalskaya, A., Gallo, M., & Camerer, C. F. (2021). The golden age of social science. *Proceedings of the National Academy of Sciences*, *118*(5), e2002923118. https://doi.org/10.1073/pnas.2002923118

Di Pietro, G. (2023). The impact of COVID-19 on student achievement: Evidence from a recent meta-analysis. *Educational Research Review*, *39*, 100530. https://doi.org/10.1016/j.edurev.2023.100530

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Halpern, D. (2015). *Inside the Nudge Unit*. Ebury Publishing.
OCLC: 1124406846.

Haushofer, J., & Fehr, E. (2014). On the psychology of poverty. *Science*, *344*(6186), 862–867. https://doi.org/10.1126/science.1232491

Hershcovits, H., Vilenchik, D., & Gal, K. (2020). Modeling engagement in self-directed learning systems using principal component analysis. *IEEE Transactions on Learning Technologies*, *13*(1), 164–171. https://doi.org/10.1109/tlt.2019.2922902

Lippens, L., Vermeiren, S., & Baert, S. (2023). The state of hiring discrimination: A meta-analysis of (almost) all recent correspondence experiments. *European Economic Review*, *151*, 104315. https://doi.org/10.1016/j.euroecorev.2022.104315

Meeter, M. (2021). Primary school mathematics during the COVID-19 pandemic: No evidence of learning gaps in adaptive practicing results. *Trends in Neuroscience and Education*, *25*, 100163. https://doi.org/10.1016/j.tine.2021.100163

Miller, G. A. (2003). The cognitive revolution: A historical perspective. *Trends in Cognitive Sciences*, *7*(3), 141–144. https://doi.org/10.1016/S1364-6613(03)00029-9

Morrison, J., Wolf, B., Ross, S., Risman, K., & McLemore, C. (2019, April). *Efficacy study of Zearn math in a large urban school district*. Center for Research and Reform in Education. http://jhir.library.jhu.edu/handle/1774.2/62395

National Academy of Sciences Staff, National Academy of Engineering, & Institute of Medicine. (2005). *Facilitating interdisciplinary research*. The National Academies Press. Retrieved August 14, 2024, from https://doi.org/10.17226/11153

OECD. (2017, March 1). *Behavioural Insights and Public Policy: Lessons from Around the World*. https://doi.org/10.1787/9789264270480-en

Pernu, T. K., & Elzein, N. (2020). From neuroscience to law: Bridging the gap. *Frontiers in Psychology*, *11*, 1862. https://doi.org/10.3389/fpsyg.2020.01862

Qiu, F., Zhang, G., Sheng, X., Jiang, L., Zhu, L., Xiang, Q., Jiang, B., & Chen, P.-k. (2022). Predicting students' performance in e-learning using learning process and behaviour data. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-021-03867-8

Quillian, L., & Lee, J. J. (2023). Trends in racial and ethnic discrimination in hiring in six western countries. *Proceedings of the National Academy of Sciences*, *120*(6), e2212875120. https://doi.org/10.1073/pnas.2212875120

Ran, H., Kasli, M., & Secada, W. G. (2020). A meta-analysis on computer technology intervention effects on mathematics achievement for low-performing students in K-12 classrooms. *Journal of Educational Computing Research*, *59*(1), 119–153. https://doi.org/10.1177/0735633120952063

Salazar, A., Serrano, A., & Vergara, L. (2007). Non-parametric ICA reveals learning styles in education activities through the web. *2007 IEEE Workshop on Machine Learning for Signal Processing*. https://doi.org/10.1109/mlsp.2007.4414316

Shin, D., & Shim, J. (2020). A systematic review on data mining for mathematics and science education. *International Journal of Science and Mathematics Education*, *19*(4), 639–659. https://doi.org/10.1007/s10763-020-10085-7

Stiglitz, J. E., Fitoussi, J.-P., & Durand, M. (2019, November 19). *Measuring what counts*. The New Press.

Thaler, R. H., & Sunstein, C. R. (2003). Libertarian Paternalism. *American Economic Review*, *93*(2), 175–179. https://doi.org/10.1257/000282803321947001

Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness* (Rev. and expanded ed). Penguin Books.

*Chapter 2*

# UNVEILING ADAPTIVE PEDAGOGY: REINFORCEMENT LEARNING MODELS ILLUMINATE TEACHER DECISION-MAKING IN AN ONLINE MATH-TEACHING PLATFORM

## 2.1 Abstract

Reinforcement learning (RL) models are widely used in computational psychology and neuroscience to understand decision-making processes in lab-controlled contexts. In this study, we offer a novel application of the Q-learning model on a large-scale digital mathematics education dataset. Analyzing data from 1,832 classrooms across an academic year, we demonstrate that Q-learning outperforms traditional logistic regression in capturing teachers' adapti ve behaviors. We further use estimated parameters from our computational models to identify teacher behavioral profiles. Our findings reveal that those with higher learning rates achieve superior student outcomes. We also observe that soci oeconomic factors correlate with model parameters, indicating potential systemic disparities in educational approaches. This approach provides new paths for understanding pedagogical strategies and generating potential interventions to enhance student educational outcomes. By demonstrating the relevance of an RL model in a digital educational environment, our study introduces a new framework for applying models traditionally used in lab settings to complex, real-world data.

## 2.2 Introduction

Predicting repeated behavior has been a long-standing goal of the behavioral sciences, including economics, psychology, and neuroscience (Hagger et al., 2023; Venkatesh et al., 2023; Verplanken & Orbell, 2022). Reinforcement learning (RL) algorithms have emerged as a prominent way of quantifying these relationships, assigning a mathematical relationship between contextual cues (states), behavior (actions), and reward (Kaelbling et al., 1996; Sutton & Barto, 2018). These algorithms have found wide application in neuroscience and cognitive psychology, where they are used in data sets to model agents in specific environments (Zhang et al., 2020). Researchers have also attempted to bridge this computational approach and clinical applications.

For example, Brown et al. (2021) investigated whether depression symptoms are related to features of reinforcement learning, fitting a state-free q-learning algorithm to participants' behavior in a reward and loss learning task (i.e., learning a stimulus's expected monetary gains or losses, respectively). The researchers then regressed depression symptoms on the model-derived parameters and found that depression symptoms may selectively disrupt specific components of the reinforcement learning process. Notably, the Brown et al. study bridges computational neuroscience and clinical applications, offering a novel way of characterizing depression, typifying patients, and opening new avenues for personalized treatment. The study resonates with using estimated parameters as markers of individual differences, hinting at the potential for personalized interventions based on these models. Similarly, our current study examines the variation in estimated learning rate parameters and their association with overall student outcomes.

Moreover, Niv et al. (2022) proposed a perspective on cognitive behavioral therapy (CBT) by framing it within reinforcement learning (RL) theory. Highlighting the similarities between these two domains offers a promising avenue for advancing our understanding of CBT and refining its clinical applications. Specifically, the researchers propose that CBT's cognitive aspect may correspond to the model-based learning system, while its behavioral component relates to the model-free system. There are two examples of RL applications. First, prolonged exposure therapy (i.e., vividly recounting a traumatic experience in a safe environment) is akin to updating the value of a traumatic state. One novel insight from RL is that moderate prediction errors (i.e., gradual extinction) are more effective than high prediction errors, which can lead to new values being assigned to an entirely different state. Second, RL theory can explain how CBT treatments for obsessive-compulsive disorder work. Patients are asked to imagine and write down the worst-case outcomes of not performing their obsessions. In RL terms, this activity forces patients to create a model of the world with low transition probabilities for worst-case scenarios, effectively reducing the perceived probability of feared outcomes. While this theoretical framework is compelling, future research must substantiate these novel insights with concrete empirical evidence.

More pragmatically, Park et al. (2019) explored the benefits of using a personalized social robot learning companion to improve engagement and learning outcomes among young English language learners. Their approach employed a model-free Q-learning strategy, incorporating affective elements such as level of engagement

and emotional states, to train a customized storytelling approach for each child. The experiment included 67 children between the ages of 4-6 divided into one of three groups: personalized robot, non-personalization robot, and a no-robot baseline. Throughout 6-8 sessions, the children participated in storytelling activities with the robots: the personalized robot used this reinforcement learning approach to predict the complexity levels of the activities to maximize a child's future engagement and learning gains. Specifically, a) the reward was a combination of engagement and learning progress (assessed through the child's use of new words and syntax structures), b) states were the users' engagement (measured by verbal and nonverbal cues) and affective arousal levels (measured by facial expressions), and c) actions were the complexity level of the activities. Results indicated that the personalized robot effectively adapted to each child's needs, leading to better engagement and learning outcomes than non-personalized and no-robot conditions. The positive results suggest that this Q-learning strategy can offer tailored and improved teaching methods. However, Park et al. made several arbitrary modeling choices without clear empirical or theoretical justification (e.g., the decreasing learning rate had a lower bound of 0.125), rather than exploring whether other parameter values yielded better outcomes. Therefore, future models could enhance the robot's outcomes through parameter optimization. Additionally, the generalizability of these findings to other educational settings and populations is needed. This approach effectively operationalizes states, actions, and rewards in an educational space, resulting in adaptive, engaging learning experiences for students.

The focus on neuroscience and cognitive psychology applications presents an opportunity to use methods from one set of disciplines to conduct research on data traditionally used in another. In this paper, we aim to further this integration by applying RL algorithms to model the decision-making process of teachers in the math-teaching platform Zearn. RL provides a system of rewards and punishments where the agent (in this case, the teacher) learns to make optimal decisions by maximizing the rewards and minimizing the punishments (Kaelbling et al., 1996; Sutton & Barto, 2018). By modeling teachers as agents in an RL framework, we assume they make decisions to maximize their cumulative rewards over time. For instance, the teacher chooses which pedagogical actions to employ, such as assigning homework, checking student progress, or reviewing content, in anticipation of enhancing student achievement. Applying RL algorithms allows for flexibility in learning the best strategy given certain contextual information. We provide a model of how teachers adapt their strategies in response to student performance and other

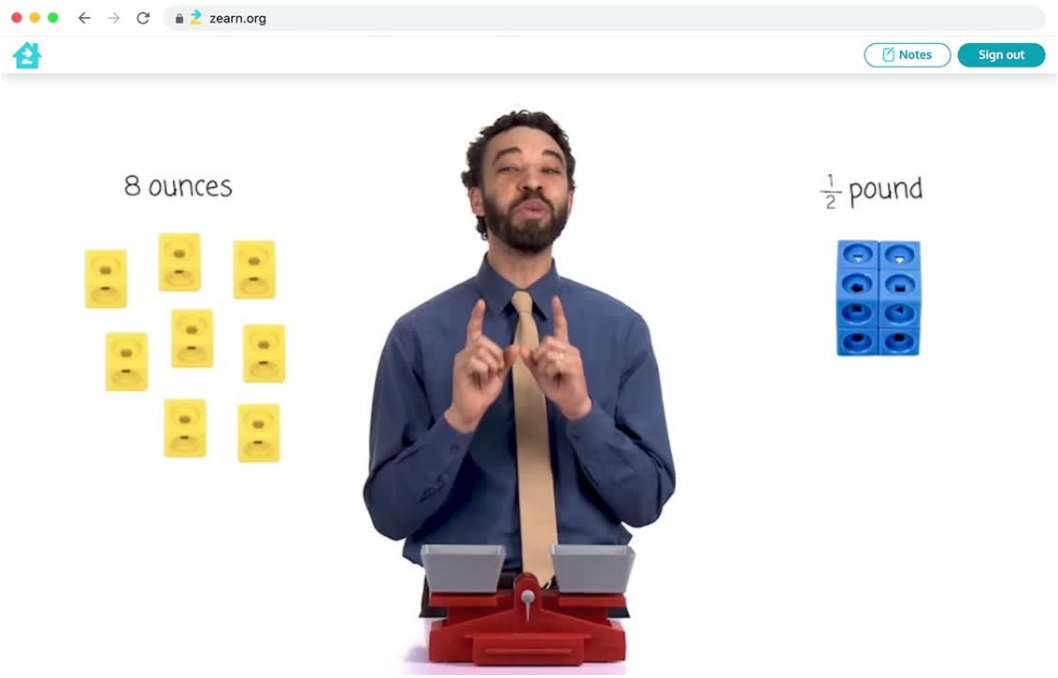


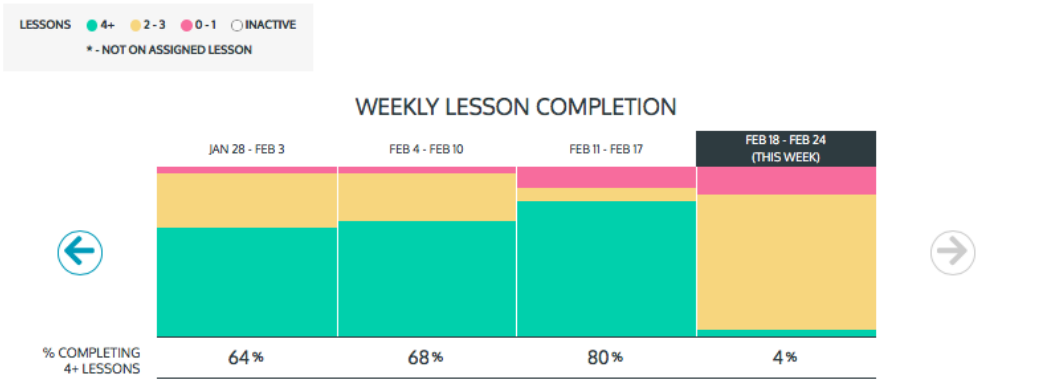

Figure 2.1: Student lesson on Zearn.
The image displays an example of an online lesson with visual models on the platform (Zearn, 2024c).

contextual factors. This approach offers a flexible model for our available data and opens new avenues for understanding and enhancing human behavior in complex, real-world settings.

**The Zearn Platform**

Zearn is a digital platform for mathematics education designed to facilitate the teaching and learning of mathematics. About 25% of elementary school students and over 1 million middle school students across the United States use Zearn (Zearn, 2024b). Its unique blend of hands-on teaching and immersive digital learning, paired with its widespread adoption, provides a promising setting for understanding how teachers adapt their strategies to optimize student achievement.

Zearn's pedagogical approach includes interactive digital lessons using visual aids (see Figure 2.1) and real-time student feedback. Students go through a series of representations — concrete, pictorial, and abstract — each designed to scaffold their understanding, i.e., "breaking down" problems (Jumaat & Tasir, 2014; Reiser & Tabak, 2014, see) and prepare them for subsequent levels.




| | | LESSONS PER WEEK | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| FIRST | LAST | JAN 28 - FEB 3 | FEB 4 - FEB 10 | FEB 11 - FEB 17 | FEB 18 - FEB 24 (THIS WEEK) | ASSIGNED LESSON | TOTAL BADGES | TOTAL BONUSES | LAST LOGIN |
| Trina | Brown | 1 | 3 | 1 | 2 | G4 M1 L10 | 10 | 8 | Wed, Feb 20, 2019, 9:45 AM |
| Paul | Derrick | 4 | 4 | 6 | 3 | G4 M2 L5 | 22 | 7 | Wed, Feb 20, 2019, 9:19 AM |
| Sebastian | Diaz | 3 | 5 | 5 | 3 | G4 M3 L2 | 24 | 15 | Wed, Feb 20, 2019, 9:03 AM |
| Janiya | DuBuque | 5 | 6 | 5 | 2 | G4 M3 L1 | 23 | 9 | Wed, Feb 20, 2019, 9:16 AM |
| Fernando | Garcia | 2 | 2 | 2 | 1 | G4 M1 L11 | 10 | 0 | Wed, Feb 20, 2019, 9:52 AM |
| Jordan | Goldstein | 4 | 5 | 5 | 3 | G4 M3 L1 | 23 | 13 | Wed, Feb 20, 2019, 9:09 AM |
| Roxanne | Jacobson | 3 | 1 | 1 | 1 | G4 M1 L9 | 8 | 6 | Tue, Feb 19, 2019, 9:08 AM |

Figure 2.2: Sample classroom report.
The image displays a summary dashboard where teachers can follow their students' progress as indicated by the number of lessons completed (Zearn, 2024d).

The platform's structure provides students with a personalized learning experience (see Figure A.2 for a screenshot of the student portal) and teachers with resources to track student progress and make informed decisions (see Figure 2.2 for a sample class report). Zearn follows a rotation model of learning—that is, a blend of traditional face-to-face learning (i.e., small group instruction) with online learning (i.e., self-paced online lessons). With this approach, students can learn new grade-level content in two distinct ways: independently, by engaging in digital lessons, and in small groups with their teacher and peers (Zearn, 2024e, 2024g).

A key feature of Zearn is its badge system, which tracks student progress and motivates continued learning (see Figure 2.3). Students earn badges upon mastery of specific skills, providing a tangible representation of their achievement. This system motivates students and provides teachers with valuable data on student performance, informing their decision-making process (Knudsen et al., 2020). Zearn also incorporates notifications, known as Tower Alerts, sent to teachers when a

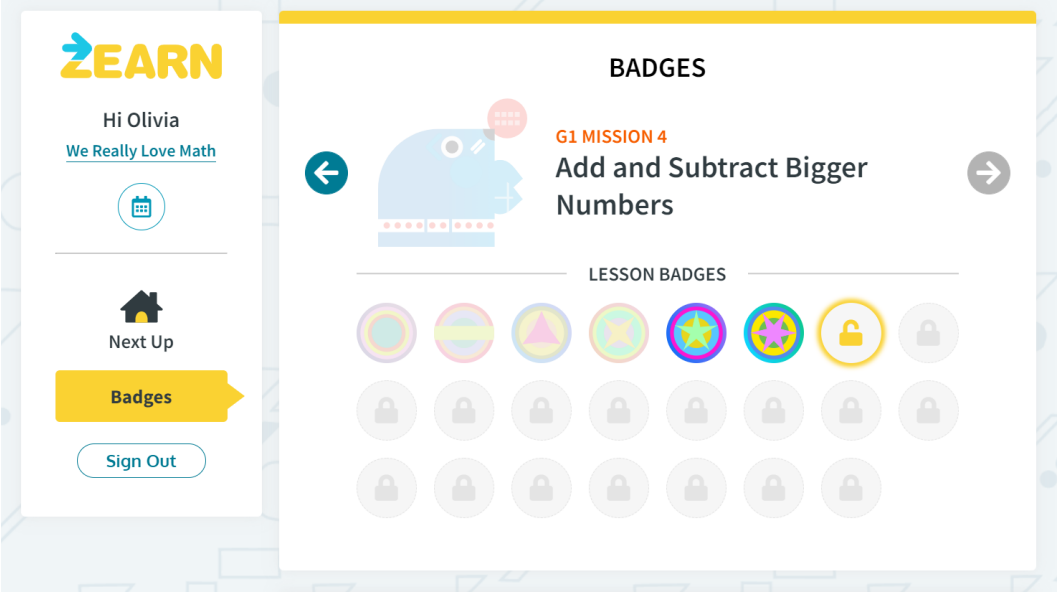



Figure 2.3: Student badges page.
The image displays a summary dashboard where students can see their total badges earned (i.e., lessons completed) for a given mission (i.e., course module). Faded badges on the image signify open lessons to be completed, unfaded badges represent earned badges, and the locked icons correspond to future digital lessons that will open once all activities in the current lesson are completed (Zearn, 2024h).

student struggles with a specific concept. This feature allows teachers to provide timely support and address learning gaps, enhancing the platform's capacity for personalized learning (Zearn, 2024i).

Morrison et al. (2019) evaluated the effectiveness of Zearn Math in a large urban school district, employing a mixed-methods approach. The study revealed mixed results. Quantitatively, no statistically significant differences in mathematics achievement gains were found between treatment and comparison students on either the Northwest Evaluation Association Measurement of Academic Progress (NWEA MAP) or state assessments. However, usage data analysis showed positive correlations between Zearn Math engagement and student outcomes. For instance, each additional hour per year using Zearn was associated with a 0.0375 point increase in NWEA MAP scores ($p < .001$), or a 0.02% increase. On the state assessment, each additional lesson completed correlated with a 0.004 standard deviation improvement ($p < .001$). Qualitatively, 83.6% of teachers agreed that Zearn Math engaged students in mathematics education. Approximately 70% of teachers reported that both digital lessons and small-group lessons promoted higher-order thinking skills. However,

only 63.3% of teachers agreed that the curriculum was effective for increasing student achievement. Notably, just under 50% of teachers felt adequately prepared to implement Zearn Math, highlighting a key challenge. The study identified specific strengths and weaknesses. The half-class rotational model was cited as a major strength, allowing for differentiated instruction and independent learning. However, this model also presented initial challenges, with teachers and administrators reporting a period of adjustment to the new instructional approach. Additionally, while 89.2% of students agreed that Zearn Math was good for learning mathematics, comparison group students showed significantly higher mathematics self-efficacy and interest. These findings underscore both the potential of digital math platforms like Zearn Math and the importance of comprehensive teacher preparation and support for effective implementation.

Another noteworthy feature is the platform's comprehensive professional development component, which is accessible to schools with a paid account (see Figure A.3 for a sample training schedule). In this program, teachers within a school collaborate to explore each unit or mission through word problems, fluencies, and small group lessons. They also analyze student work and problem-solving strategies. This professional development prioritizes (1) each mission's primary mathematical concept, (2) visual representations to scaffold learning, and (3) strategies to address unfinished learning from prior grades while preparing for future learning (Morrison et al., 2019).

Researchers have also examined the Zearn approach for teachers' professional development. For example, Knudsen et al. (2020) focused on the effectiveness of the Curriculum Study Professional Development (CS PD) program developed by Zearn to enhance elementary school teachers' Pedagogical Content Knowledge (PCK) in teaching mathematics. The researchers used a case study approach, examining eight teachers across various schools and districts who had undergone the CS PD program. Data collection methods included think-aloud interviews, classroom and CS PD session observations, and interviews with teachers and administrators, and the researchers found that 75% of the teachers experienced growth in their PCK. Knudsen et al. (2020) found that the key strengths of the CS PD program included its relevance to practice, encouragement of collaboration among teachers, and its effectiveness in developing big ideas. However, challenges in responding to students' in-the-moment problem-solving and varied teacher engagement were noted.

The findings underscore the importance of enhancing teachers' pedagogical content knowledge, especially in mathematics.

One of the measures we developed in our study specifically captures pedagogical content knowledge (see section 2.4). These insights can inform the development of similar professional development programs in the current project, ensuring they are effectively tailored to teachers' instructional needs and students' diverse backgrounds. Further, we are able to model teacher behavior at the individual level.

Zearn's integrated framework provides a rich repository of data for our analysis. The variables delineated for investigation by Zearn encompass: (1) teacher engagement, quantified through a diverse set of actions (see Materials and Methods and Table A.1; (2) student achievement, denoted by variables such as lesson completion (i.e., *badges* earned after each lesson is finished with full proficiency); and (3) student struggles, monitored through variables such as *tower alerts* (see Table A.1 for a full glossary of available variables).

**Research Questions**

We proposed the following research questions:

1. Characterizing Teacher Behavior: How can we best explain teachers' action choices? How does the explanatory power of reinforcement learning compare to simpler baseline models? Which specific reinforcement learning model best captures the empirical data on teacher behavior?

2. Impact of Estimated RL Parameters: How do individual differences in teachers' decision-making patterns, as inferred from the parameters of the best-fitting reinforcement learning model, relate to heterogeneity in student achievement gains? Can we identify specific teacher behavioral profiles that predict better student learning outcomes?

3. Influence of Teacher and School Background: To what extent do school contextual factors (e.g., socioeconomic status) account for variation in teachers' instructional choices, as quantified by the parameters of the reinforcement learning model?

**Hypotheses**

The hypotheses are:

1. The Q-learning model will effectively capture key behavioral signatures of teacher learning and decision-making, specifically: (a) teachers will more likely

select actions with higher Q-values, (b) average prediction errors will decrease over time, and (c) Q-values will converge toward empirical reward rates over time.

2. The Q-learning model will outperform traditional methods (such as logistic regression) in explaining teachers' adaptive strategies, as measured by the Akaike information criterion (AIC).

3. Individual Q-learning model parameters (e.g., learning rate) will be significantly associated with measures of student performance and contextual factors.

### 2.3 Theory

**Reinforcement Learning to Capture Patterns in Repeated Behavior**

In RL, an agent learns to make decisions over time. Formally, an RL task is a tuple $\langle S, A, R, P, \gamma \rangle$, where:

- $S$ is a set of states, that is, the possible configurations or situations in which the agent can find itself. The agent must be able to perceive the state of its environment to some extent. In cases with multiple states, it must also be able to take actions that affect, or change, the state (Sutton & Barto, 2018, p. 2).
- $A$ is a set of actions, that is, choices available to the agent.
- $R_t = \mathbb{E}[R_{t+1}|S_t = s, A_t = a]$ is a reward function (Sutton & Barto, 2018, p. xx). This is a signal from the environment to the agent and corresponds to the purpose or goal of the agent (Sutton & Barto, 2018, p. 53).
- $P = \Pr[S_{t+1} = s'|S_t = s, A_t = a]$ is a state transition probability, a function that describes the probability of moving to state $s'$, given the current state $s$ and action $a$.
- $\gamma \in [0, 1]$ is a discount factor, a value that determines how much the agent values future rewards compared to immediate ones.

We also define the agent's decisions as a probability distribution over actions, namely, the policy $\pi(a|s) = \Pr[A_t = a|S_t = s]$, which maps "perceived states of the environment to actions to be taken when in those states" (Sutton & Barto, 2018, p. 6).

RL models have been used to explain learning and reward association (Rescorla & Wagner, 1972; Thorndike, 1931). One common approach in human studies is

to apply the *multi-armed bandit* task (Daw et al., 2006; Dennison et al., 2022). In this type of experiment, participants are presented with multiple actions, each with an unknown payoff. The subject's goal is to learn the best outcome through trial and error. In the beginning, the reward-action relationships are unknown, so the participant must explore or sample each action (Sutton & Barto, 2018, p. 3). This exploration-exploitation trade-off (i.e., trying new actions to learn about their associated rewards versus sticking with actions known to yield good rewards) is a central theme in RL (Daw et al., 2006; Wyatt et al., 2024) and has the potential to provide valuable insights into how individuals learn and make decisions over time.

In the context of education and teaching, RL has been present as early as 1960, with Ronald Howard applying this mathematical framework to instruction theory (Howard, 1960). Later, in 1972, Richard Atkinson proposed a theory of instruction that encapsulates the key components of a Markov decision process, including states, actions, transition probabilities, reward functions, and a time horizon (Atkinson, 1972). In Atkinson's framework, actions are instructional activities (e.g., assigning problem sets) that can change a given state (e.g., student learning level). These changes in states can yield rewards minus the associated cost of the action. For example, a teacher may be rewarded with an increase in the knowledge or skill of a student, but such reward must be balanced with its associated effort (e.g., labor cost). Atkinson and colleagues continued to test many parameterizations of this idea, contributing significantly to the development of RL theory in the context of education (see Doroudi et al., 2019; Memarian & Doleck, 2024, for full reviews.).

While RL has been applied in educational settings, our work represents a significant departure from the existing literature. For instance, Memarian and Doleck (2024) highlight that the vast majority of RL application in education focus on automated systems and artificial agents, with little work on improving teaching and learning processes. This work addresses this gap by applying RL principles to human teacher decision-making. We aim to demonstrate ways in which RL can inform and optimize instructional choices by Zearn teachers. Furthermore, unlike the much of the literature, with models that seek to maximize student outcomes, we also aim to ensure our findings uncover meaningful behavioral patterns already present in the data (i.e., through model falsification, see Palminteri et al., 2017, and section 2.4).

In the Zearn context, we define the decision process as follows:

1. Agents are the teachers.

2. Actions include the teachers' choices of specific pedagogical strategies.
3. The reward is a function of the average student performance or activity within a classroom.

Note that the models presented here do not use a set of states $S$. Further, we only consider model-free RL algorithms, which are so called because they do not require the agent to learn $P$ (i.e., state transition probabilities) to approximate expected rewards (Watkins & Dayan, 1992). In the following sections, we describe the Q-learning model we use to capture teacher behavior in the Zearn platform.

**Q-Learning Model**

Consider a teacher using the Zearn platform. Each week, they must decide between assigning additional homework (action 1) or not assigning additional homework but spending their time on another task (the outside option, or action 0). At first, the teacher is uncertain about the best action to take. They start with initial beliefs about each action's long-term value (Q-value). However, they know that these beliefs may not be accurate and that they need to learn from experience.

Each week, the teacher chooses an action based on the current Q-value estimates. For example, in week 1 the teacher believes that assigning homework (action 1) has a slightly higher Q-value than other available activities (action 0). So, they assign homework and observe the outcome.

Then, the teacher receives a reward signal (e.g., the students' performance after the homework assignment). They use this reward to update their estimate of the Q-value for assigning homework, following the Q-learning update rule. This rule adjusts the Q-value estimate based on the difference between the observed reward and the previous estimate multiplied by a learning rate parameter.

Over the following weeks, the teacher continues to make decisions and update their estimates based on the outcomes observed. Sometimes, they explore new actions to gather more information, even if these actions do not seem optimal based on the current estimates. Other times, the teacher exploits their experience by choosing the action with the highest estimated Q-value. The degree of exploration (vs. exploitation) is controlled by a temperature parameter (i.e., $\tau$).

As the teacher learns from experience, the Q-value estimates gradually converge toward the true values for each action. We outline sample values for this teacher in table 2.1.

| **Week** | **Q-value Difference** | **Policy** | **Choice** | **Reward** | **Prediction Error** | **Updated Q-value Difference** |
|---|---|---|---|---|---|---|
| $t$ | $\Delta Q_t$ | $\Pr_t(a = 1)$ | $a$ | $R_t$ | $\delta_t$ | $\Delta Q_{t+1}$ |
| 1 | 1.533 | 0 | 0 | 0.052 | -2.891 | 0.815 |
| 2 | 0.815 | 0 | 0 | 0.163 | -2.122 | 0.288 |
| 3 | 0.288 | 0.045 | 0 | 0.047 | -1.649 | -0.121 |
| 4 | -0.121 | 0.782 | 1 | 0.039 | -0.018 | -0.125 |
| 5 | -0.125 | 0.790 | 1 | 0.064 | -0.030 | -0.133 |
| ... | | | | | | |

Table 2.1: Example of a Q-learning algorithm for a teacher on the Zearn platform. Each week ($t$), the teacher decides between Choice = 1 (assigning homework) and Choice = 0 (not assigning homework) based on the difference in Q-values ($Q_t$) for each action. The teacher's choice ($a$) is determined by the policy ($\Pr_t(a = 1) = 1/(1 + e^{-\tau \Delta U_t(a)})$, where $\Delta U_t = \Delta Q_t - \text{cost}(a = 1)$). After observing the reward ($R_t$) associated with the chosen action, the teacher computes the prediction error ($\delta_t$) using 2.4. The Q-value difference is then updated using by $\Delta Q_t(a) = \Delta Q_{t-1}(a) + \alpha \delta_t$. As the teacher learns from experience, the Q-value difference converges toward the highest reward. The values used here were drawn from the estimate of an actual Zearn account ($\alpha = 0.25$, $\tau = 10.57$, $\gamma = 0.46$, cost = 1.38).

This model frames decision-making as a result of accumulated experience and the anticipation of future rewards. In other words, Q-learning involves the iterative refinement of Q-value functions, which map an agent's actions to evolving expectations of future rewards (analogous to subjective value or utility). This methodological approach is closely related to the classic multi-armed bandit problem, wherein the agent faces a finite set of choices (e.g., slot machines), each linked to a specific reward schedule, and aims to learn the action that yields the highest returns. Learning in this model depends on adjusting expectations to reduce the impact of prediction errors (the "surprise level," or the difference between expected and realized outcomes).

In this study, we opt for a state-independent version of Q-learning. That is, the Q-values do not vary with a contextual variable but are learned for each action only. This assumption is useful in scenarios where the state exerts minimal influence on the outcome of the action or when the state is difficult to define or observe (Sutton & Barto, 2018). As such, the Q-function represents the expected return or future reward for taking action $a \in A = \{a_1, a_2, ...\}$ following a certain policy $\pi = \Pr(a)$. The updating rule uses the equation:

$$Q_t(a) = Q_{t-1}(a) + \alpha\delta_t \tag{2.1}$$

where $\alpha$ is the learning rate, which determines how much the Q-value is updated based on $\delta$, the reward prediction error. The reward prediction error is the difference between the estimated Q-value and the observed discounted reward. This error is used to update the Q-value of the chosen action in the direction of the observed reward, scaled by the learning rate $\alpha$, as follows:

$$\delta_t = \gamma R_t(a) - Q_{t-1}(a) \tag{2.2}$$

where:

- $a$ is the chosen action,
- $R_t(a)$ is the immediate reward received after taking action $a$,
- $\gamma$ is the discount factor[1],
- $Q_{t-1}(a)$ is the estimate of the Q-value for action $a$ in the previous period.

In other words, $\alpha$ is the extent to which the newly acquired information will override the old information. A value of 0 means the agent does not learn anything. The agent starts with an initial Q-value and then updates the Q-values based on the experiences it gathers from interactions with the environment. The update rule is applied every time the agent takes action $a$ and receives a reward $R$. The agent selects actions based on a policy function of the Q-values. A common choice is the softmax action selection method, which chooses actions probabilistically based on their Q-values, as follows:

$$\Pr_t(a) = \frac{e^{\tau U_t(a)}}{\sum_{a'} e^{\tau U_t(a')}} \tag{2.3}$$

where:

- $\Pr_t(a)$ is the probability of choosing action $a$ at time $t$,

[1] Commonly, this parameter captures the degree to which future rewards are discounted compared to immediate rewards. In this example, it could also act as a scaling factor of net reward.

- $U_t(a) = Q_t(a) - \text{cost}(a)$ is the utility of action $a$ at time $t$, which is the difference between the Q-value of action $a$ ($Q_t(a)$) and the cost associated with taking that action,
- $\tau$ is a parameter known as the inverse temperature, or the degree of randomness in the choice behavior,[2]
- $\text{cost}(a)$ is the perceived effort or inconvenience associated with action $a$,
- The denominator is the sum over all possible actions $a' \in A$ of the exponential of their Q-values multiplied by the inverse temperature, and it functions as a normalizing value.

**Binary Actions**

For cases in which actions are binary (i.e., only two options $a = 0$ and $a = 1$), we can simplify the general Q-learning framework by tracking the value of one action relative to an outside option with a Q-value and cost of zero (i.e., base value $Q(a = 0) = 0$ and $\text{cost}(a = 0) = 0$). Throughout our analyses, the action represents engaging with the platform (see Results for details), while the outside option represents not engaging.

Rather than tracking separate Q-values for each action, we track their difference $\Delta Q_t = Q_t(a = 1) - Q_t(a = 0)$. When action $a = 1$ is chosen, we update this difference based on its reward relative to the baseline Q-value difference. When action $a = 0$ is chosen and yields a reward, we need to decrease $\Delta Q_t$ to indicate that the outside option is more valuable than previously estimated. We start from the update equations for each action and our definition of $\Delta Q_t$:

$$
\begin{aligned}
Q_t(a = 1) &= Q_{t-1}(a = 1) + \alpha(\gamma R_t - Q_{t-1}(a = 1)) \\
Q_t(a = 0) &= Q_{t-1}(a = 0) + \alpha(\gamma R_t - Q_{t-1}(a = 0)) \\
\Delta Q_t &= Q_t(a = 1) - Q_t(a = 0).
\end{aligned}
$$

[2] One possible interpretation of the inverse temperature parameter $\tau$ is the agent's confidence in its Q-values, which controls the trade-off between exploration and exploitation. When $\tau$ is high, the agent explores more because the action probabilities are more uniform. When $\tau$ is low, the agent exploits more because the action with the highest Q-value is more likely to be chosen than the others. Some models may also allow for agents to start with a high inverse temperature to encourage exploration and then gradually decrease it to favor the exploitation of the learned policy.

Then we derive the update rule for the difference in Q-values. When $a_{t-1} = 1$, we update only $Q_{t-1}(a = 1)$:

$$\begin{aligned}\Delta Q_t &= Q_{t-1}(a = 1) + \alpha(\gamma R_t - Q_{t-1}(a = 1)) - Q_{t-1}(a = 0) \\ &= \Delta Q_{t-1} + \alpha(\gamma R_t - Q_{t-1}(a = 1)) \\ &= \Delta Q_{t-1} + \alpha(\gamma R_t - \Delta Q_{t-1}) \quad \text{since } Q(a = 0) = 0,\end{aligned}$$

and vice versa when $a_{t-1} = 0$:

$$\begin{aligned}\Delta Q_t &= Q_{t-1}(a = 1) - (Q_{t-1}(a = 0) + \alpha(\gamma R_t - Q_{t-1}(a = 0))) \\ &= \Delta Q_{t-1} - \alpha(\gamma R_t - Q_{t-1}(a = 0)) \\ &= \Delta Q_{t-1} - \alpha(\gamma R_t) \quad \text{since } Q(a = 0) = 0 \\ &= \Delta Q_{t-1} + \alpha(-\gamma R_t).\end{aligned}$$

Thus, $\delta_t$, the prediction error for $\Delta Q_t$, can be written as:

$$\delta_t = \begin{cases} \gamma R_t - \Delta Q_{t-1} & \text{if } a = 1 \\ -\gamma R_t & \text{if } a = 0 \end{cases} \tag{2.4}$$

where:

- $\Delta Q_t = Q_t(a = 1) - Q_t(a = 0)$ is the estimate of the difference in Q-values for the two actions,
- $\alpha$ is the learning rate,
- $R_t$ is the immediate reward received after taking the chosen action.

Furthermore, the probability of choosing a particular action is determined by the logistic function as follows:

$$\begin{aligned}\Pr_t(a = 1) &= \frac{1}{1 + e^{-\tau \Delta U_t}} \\ \Pr_t(a = 0) &= 1 - \Pr_t(a = 1)\end{aligned}$$

with utility difference:

$$\begin{aligned}\Delta U_t &= Q_t(a = 1) - \text{cost}(a = 1) - (Q_t(a = 0) - \text{cost}(a = 0)) \\ &= \Delta Q_t - \text{cost}(a = 1) \quad \text{since cost}(a = 0) = 0.\end{aligned}$$

**Why Reinforcement Learning?**

Reinforcement Learning (RL) presents a few advantages over models that employ a static approach to link teacher efforts with student outcomes. It embodies the flexibility to adapt and evolve strategies over time. This dynamic framework reflects the continuous learning process seen in biological systems and offers a way to model the ongoing adaptations that occur in teacher-student interactions in the classroom. Beyond our immediate study goals, RL models hold the potential for automating instructional decisions based on identified patterns, potentially alleviating the workload on teachers and optimizing the educational process.

In this study, we treat teachers as agents who navigate their environment (i.e., the classroom) by taking actions based on their observations and the feedback they receive. RL algorithms can characterize individual profiles for teachers, providing insights into how they adapt and respond to various states and rewards within the educational setting. By estimating individual teacher parameters, RL provides insights into valuable aspects for policymakers to design targeted interventions aimed at enhancing educational outcomes.

Further, the flexibility of RL makes it an ideal tool to model how teachers address changing classroom needs. By incorporating a wide range of variables (e.g., actions and rewards), RL models are customizable to diverse educational contexts and objectives. Given this flexibility in mathematically mapping the agent-environment interaction (i.e., many models potentially satisfy our initial assumptions), our first step is a competition of models, selecting a set of models applicable to our setting, fitting them to the data, and comparing their performances.

## 2.4 Results

**Selecting a model specification**

To analyze Zearn data spanning an entire academic year, we first establish a framework for actions and rewards. In this framework, teacher activities drive the educational process, and student achievements result from these efforts. Instead of relying solely on one analytical approach, our strategy involves a large set of candidate models, as shown in Table 2.6. Our overarching goal was to strengthen the reliability of our findings and offer a detailed understanding of the underlying behavioral patterns.

**Dimensionality Reduction**

We first conducted a dimensionality reduction with Non-negative Matrix Factorization (NMF) and four components (see Figures A.7 and A.8 for a comparison of different methods by balancing reconstruction accuracy, i.e., R-squared, with clustering clarity, i.e., Silhouette Scores). While our dataset offers many potential action and reward variables, the direct use of these variables presents significant challenges:

1. **Complexity**: The sheer number of available variables complicates the identification of meaningful patterns and relationships.
2. **Dimensionality**: The high-dimensional nature of the data risks diluting important signals due to the "curse of dimensionality."[3]
3. **Interpretability**: Directly interpreting the impact of specific actions or behaviors on outcomes can be complicated by the intertwined nature of the data.

By reducing the data to a manageable number of components, we can more readily identify underlying patterns of behavior and interaction. To achieve this, we applied Nonnegative Matrix Factorization (NMF) and used the results to define action and reward variables rather than using individual metrics. One desirable feature of NMF is that it produces sparse components, providing a distilled representation of the data, where each one reflects a combination of behaviors or activities with a potential thematic linkage. Given that our chosen RL models require discrete action variables, we choose to split teacher actions into binary variables, following a median split. In our case, a median split is equivalent to giving a value of 1 to any positive value.

**Interpreting Components** After analyzing the NMF data, we identified four significant components for teachers and students. Figure 2.4 displays these components as heatmaps, offering insight into the underlying behavioral structures. Given the loadings, we interpret the components as follows:

[3]Richard Bellman coined this phrase to describe the challenge of optimizing a control process by searching over a discrete multidimensional grid, where the number of grid points increases exponentially with the number of dimensions. He wrote: "In view of all that we have said in the foregoing sections, the many obstacles we appear to have surmounted, what casts the pall over our victory celebration? It is the curse of dimensionality, a malediction that has plagued the scientist from the earliest days" (Bellman, 2015).

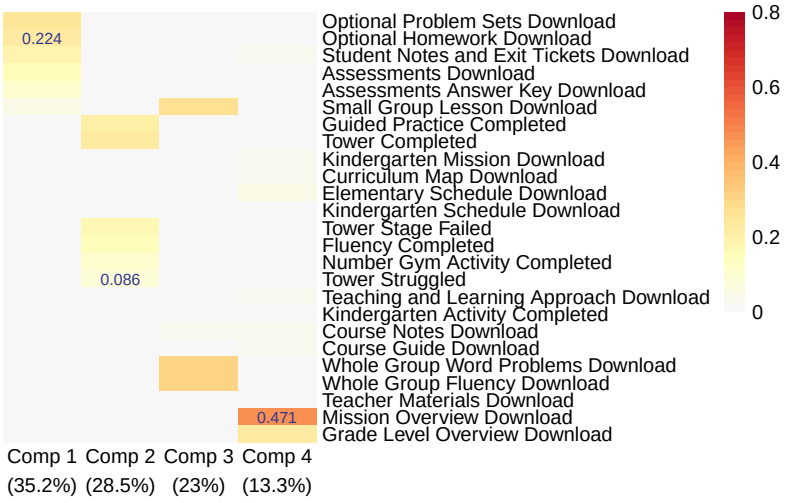
**Teachers Components**

- **Component 1 (Assessments)**: This component has substantial weights on supplemental assessment materials, such as "Optional Problem Sets Download," "Optional Homework Download," and "Student Notes and Exit Tickets Download," indicating a proactive approach to evaluating and supporting student learning progress. It could also reflect a proactive approach to monitoring student understanding and providing feedback.
- **Component 2 (Pedagogical Knowledge)**: The high weights on "Guided Practice Completed," "Tower Completed," "Tower Stage Failed," and "Fluency Completed" suggest that this component reflects when teachers are engaged in acquiring subject-matter-specific pedagogy, learning to scaffold and explain concepts in various ways.
- **Component 3 (Group Instruction)**: This component, with prominent weights on "Small Group Lesson Download," "Whole Group Word Problems Download," and "Whole Group Fluency Download," suggests a pedagogical approach focused on fostering interactive and comprehensive classroom instruction. It implies engagement in activities that promote group learning dynamics and collective problem-solving skills.
- **Component 4 (Curriculum Planning)**: The dominance of "Mission Overview Download" and "Grade Level Overview Download" in this component suggests that teachers are highly involved in strategic planning and curriculum mapping. It involves organizing the curriculum content and structuring lesson plans to align with grade-level objectives and mission overviews.

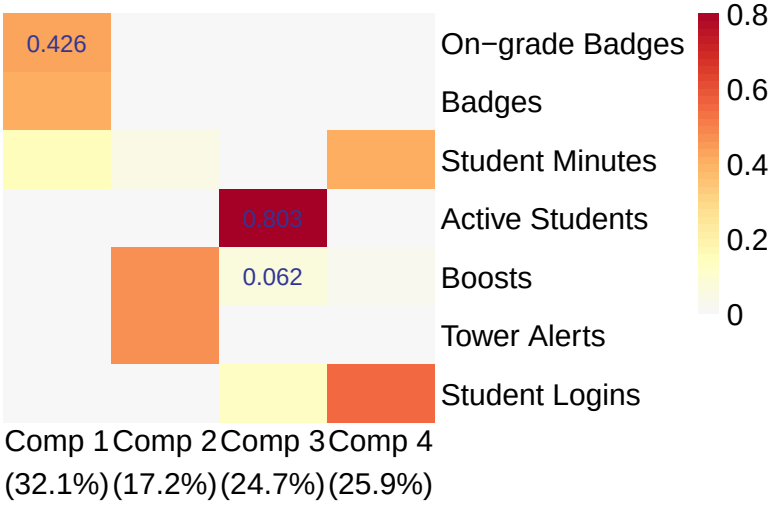
**Student Components**

- **Component 1 (Badges)**: This component emphasizes "On-grade Badges" and "Badges," indicating that it measures students' overall engagement and advancement through the curriculum.
- **Component 2 (Struggles)**: This component, which heavily weights "Boosts" and "Tower Alerts," seems to capture the frequency of occasions when students require additional scaffolding and assistance.

- **Component 3 (Number of Students)**: This component mainly consists of "Active Students," which provides insight into what proportion of students regularly log in to complete Digital Lessons.
- **Component 4 (Activity)**: Dominated by "Student Minutes" and "Student Logins," this component highlights the amount of time students invest in Zearn and the frequency of their interactions with it.




(a) Teacher data. Total reconstruction R-squared: 0.387.




(b) Student data. Total reconstruction R-squared: 0.952.

Figure 2.4: Heatmap of Non-negative Matrix Factorization (NMF) components for teacher and student data.

The rows represent the original variables, and the columns correspond to the components. The color gradient indicates the relative importance of each variable within a component based on the proportion of the component's total weight attributed to that variable. These proportions were calculated by normalizing each variable weight within a component so that they all sum to 1. The heatmaps label examples of low, moderate, and high proportion values. (a) Component 1 (Assessments) focuses on using supplemental materials for student evaluation; Component 2 (Pedagogical Knowledge) emphasizes developing subject-specific teaching strategies; Component 3 (Group Instruction) centers on collaborative and whole-class teaching methods; Component 4 (Curriculum Planning) highlights planning and lesson preparation. (b) Component 1 (Badges) measures curriculum engagement and progression; Component 2 (Struggles) indicates the need for additional academic support; Component 3 (Number of Students) tracks student participation within the platform; Component 4 (Activity) reflects the overall time spent and frequency of platform usage. The percentages in parentheses below each component label represent that component's contribution to the overall reconstruction of the data, indicating its relative importance in the NMF.

**Feature Selection**

In order to pre-select the most appropriate action and reward variables, we estimated the Q-learning models alongside a) a baseline model (constant-only regression), b) a logistic regression model inspired by dynamic analysis (Lau & Glimcher, 2005), and c) a Q-learning model, and d) a simplified Q-learning model with no cost parameter and a starting Q-value of 0. This approach acted as a filter to capture the action-reward configurations displaying the best fit. More specifically, we applied reward structures extracted from classroom data via non-negative matrix factorization (NMF) with the Frobenius Non-negative Double Singular Value Decomposition (NNDSVD) and actions derived similarly from teacher data. Then, we selected one teacher component as the action and one student component as the reward, yielding 16 configurations (4 possible actions and 4 possible rewards).

To account for temporal dynamics of actions influenced by lagged rewards, the logistic regression models incorporated lagged variables. We tested lags ranging from one to six weeks, accounting for temporal autocorrelation and potential delayed effects. The results suggest a preference for a lag of two periods as optimal (based on the "elbow" in the Area Under the Receiver Operating Characteristic curve (AUC) and the minima in the Bayesian Information Criterion (BIC) curves, see Figure A.9).

We evaluated the performance of each model configuration using log-likelihood values. Table 2.2 provides the log-likelihood scores for the four models across all 16 action-reward configurations. Our analysis revealed that, across all model types, the "Pedagogical Knowledge" action consistently showed the best fit, as evidenced by the highest (least negative) log-likelihood values. Further, the full Q-learning model consistently outperformed the others. Within Q-learning, models incorporating "Badges" and "Activity" as rewards generally outperformed others, although differences were small.

**Model Performance and Behavioral Signatures**

Given these findings, we focus our subsequent analyses on the models featuring "Pedagogical Knowledge" as the action and "Badges" as reward (see Appendix for similar analyses with "Activity" as reward). In the next stage of our analyses, we used the Akaike Information Criterion (AIC) to further compare the performance of our models, while adjusting for model complexity. Table 2.3 presents the number of parameters and mean AIC values for each model (see figure A.10 for an empirical cumulative distribution function). The scores were computed using individual

| Action | Reward | Q-learning | Q-learning (cost-free) | Lau & Glimcher | Baseline |
|---|---|---|---|---|---|
| Assessments | Badges | −24, 208.42 | −27, 425.17 | −24, 831.02 | −27, 176.43 |
| | Struggles | −24, 283.40 | −27, 709.75 | −24, 833.22 | −27, 176.43 |
| | No. Students | −24, 252.29 | −27, 108.38 | −24, 907.33 | −27, 176.43 |
| | Activity | −24, 220.46 | −27, 373.89 | −24, 827.77 | −27, 176.43 |
| Pedagogical Knowledge | Badges | −20, 900.34 | −25, 291.43 | −21, 189.95 | −23, 605.37 |
| | Struggles | −20, 900.96 | −25, 531.27 | −21, 193.80 | −23, 605.37 |
| | No. Students | −20, 906.79 | −25, 046.88 | −21, 253.18 | −23, 605.37 |
| | Activity | −20, 910.36 | −25, 213.89 | −21, 175.05 | −23, 605.37 |
| Group Instruction | Badges | −23, 788.51 | −27, 688.01 | −24, 540.78 | −27, 181.44 |
| | Struggles | −23, 835.75 | −27, 935.43 | −24, 552.94 | −27, 181.44 |
| | No. Students | −23, 800.23 | −27, 530.45 | −24, 618.32 | −27, 181.44 |
| | Activity | −23, 771.94 | −27, 591.98 | −24, 537.39 | −27, 181.44 |
| Curriculum Planning | Badges | −24, 661.42 | −28, 315.00 | −24, 900.79 | −26, 967.87 |
| | Struggles | −24, 669.08 | −28, 564.54 | −24, 903.75 | −26, 967.87 |
| | No. Students | −24, 687.01 | −28, 262.99 | −24, 980.06 | −26, 967.87 |
| | Activity | −24, 687.32 | −28, 256.58 | −24, 900.18 | −26, 967.87 |

Table 2.2: Model comparison based on log-likelihood.

The table presents log-likelihood values for different models across various action-reward combinations. Less negative values indicate better fit. All models are compared against a baseline intercept-only regression model. The cost-free Q-learning models do not include a cost parameter and assume the starting q-value to be zero. Across all models, the 'Pedagogical Knowledge' action shows the best fit overall. Note that these estimates are not hierarchical; they represent the individual estimates of each teacher. For all models, $N = 1,832$.

model-fitting with Matlab's cbm toolbox, with lower AIC values indicating better model fit. The baseline model, with its advantage in parsimony, achieved the lowest AIC, followed by the full Q-learning model. While the Q-learning model shows higher log-likelihood than the baseline in non-hierarchical estimation (see Tables 2.2 and A.4), this advantage is overwhelmed by the complexity penalty in the AIC calculation. With an average of 25.7 weeks of data per teacher, the additional parameters in the Q-learning model incur a substantial penalty relative to the log-likelihood values. Interestingly, hierarchical estimation narrows this gap, with the Q-learning model achieving a Leave-One-Out Information Criterion (LOOIC) 90.2 (SE = 34.0) points lower, suggesting improved out-of-sample prediction accuracy when parameter information is shared across teachers. (see Table 2.3).

Furthermore, the baseline model does not provide insights into the decision-making process of teachers, which is a key goal of this study. Further, Palminteri et al. (2017) stresses the importance of verifying a model's ability to produce a behavioral effect of interest. As such, we proceed with a more detailed examination of the behavioral signatures in our data and investigate which models best capture them, excluding the simplified Q-learning model, which underperformed in both measures of model fit.

| Model | Non-Hierarchical | | | Hierarchical | |
|---|---|---|---|---|---|
| | $N_{par}$ (per teacher) | LL | AIC | LL | LOOIC |
| Q-learning | 5 | $-20,900$ | $60,121$ | $-23,827$ | $50,380$ |
| Q-learning (cost-free) | 3 | $-25,291$ | $61,575$ | | |
| Lau & Glimcher | 5 | $-21,190$ | $60,700$ | | |
| Baseline | 1 | $-23,605$ | $50,875$ | $-24,478$ | $50,561$ |
| Mean number of weeks: 25.7 (SD = 5.9) | | | | | |
| Number of classrooms: 1,832 | | | | | |

Table 2.3: Summary of model fits for Pedagogical Knowledge as actions and Badges as rewards.

Non-hierarchical estimates were obtained by fitting models independently for each classroom using Matlab's cbm toolbox. Hierarchical estimates were obtained through Hamiltonian Monte Carlo sampling in Stan. In non-hierarchical fits, the Akaike Information Criterion (AIC) is calculated from the Log-Likelihood (LL) and sum of parameters across all classrooms ($-2\text{LL} + 2 \cdot 1832 \cdot N_{\text{par}}$), with lower values indicating better fit. For hierarchical fits, the Leave-One-Out Information Criterion (LOOIC) estimates out-of-sample prediction accuracy following Silva and Zanella (2024), with lower values indicating better fit. The cost-free Q-learning and Lau & Glimcher models were not estimated hierarchically due to their inferior non-hierarchical performance.

**Reward Seeking** We first examine whether teachers display a preference for choices with higher relative expected values, as measured by the estimated Q-value difference. Figure 2.5a demonstrates this reward-seeking behavior in the data ($b = 1.02, 95\%$ CI $[0.947, 1.09]$). It also shows that the non-hierarchical Q-learning model most closely fits the observed data ($b = 1.34$, 95% CI $[1.29, 1.39]$). The logit model underestimates the probability of choices with higher expected values and overestimates choices with lower ones ($b = 0.371, 95\%$ CI $[0.333, 0.409]$). The baseline model, as expected, does not capture the reward-seeking behavior and estimates a fixed choice probability ($b = 0.005, 95\%$ CI $[0.002, 0.007]$). Meanwhile, Figure 2.6a shows the hierarchical Bayesian model presents weaker but still present reward-seeking behavior ($b = 0.482, 95\%$ CI $[0.453, 0.511]$).

**Uncertainty Aversion** Next, we examine whether teachers are averse to choices with a more uncertain reward relationship. For each action, we calculated the cumulative standard deviation of rewards received when that action was chosen. The action with the higher standard deviation of rewards was designated as the uncertain option. Figure 2.5b provides insights into this behavioral signature. The x-axis represents the percentile rank of the difference in expected value (EV) between uncertain and certain options, while the y-axis shows the proportion of times teachers chose the uncertain option. The key measure in this plot is the location at which the teacher is indifferent between uncertain and certain choices. The data show uncertainty indifference at the 50.6 percentile (95% CI $[40.7, 60.6]$), and the models estimate similar indifference points (Q-learning: 56.5, 95% CI $[44.9, 68.2]$; logit: 50.2, 95% CI $[39.1, 61.3]$; baseline: 53.2, 95% CI $[40.0, 66.5]$). The hierarchical models also estimate similar indifference points (Q-learning: 52.7, 95% CI $[40.4, 65.1]$, see Figure 2.6b). Thus, this analysis cannot differentiate the three models in their ability to capture this signature.

**Evidence for Learning** Given the Q-learning model best captures reward-seeking behavior, we also examine its performance in capturing learning processes in teachers' decision-making. Figures 2.5c and 2.6c illustrate the prediction error over the course of the academic year, averaged across classrooms. This measure is an indicator of how well teachers are able to anticipate the outcomes of their actions. In the nonhierarchical model, teachers demonstrate initially large prediction errors that gradually decrease in magnitude throughout the year. On the other hand, the hierarchical estimates are comparatively smaller but still trend toward zero with

time. This consistent decrease in prediction errors suggests that teachers may be progressively improving their ability to predict the rewards associated with their choices.

Furthermore, Figures 2.5d and 2.6d provide additional insight into the learning process by showing the evolution of Q-values and reward differences over the academic year, averaged across classrooms. Here, the two estimation approaches reveal distinct patterns. The non-hierarchical model demonstrates a gradual convergence between initially high Q-value differences and the relatively stable empirical difference in mean rewards (from the previous four weeks). The hierarchical model, however, shows Q-values and reward differences tracking much more closely for a large part of the school year.

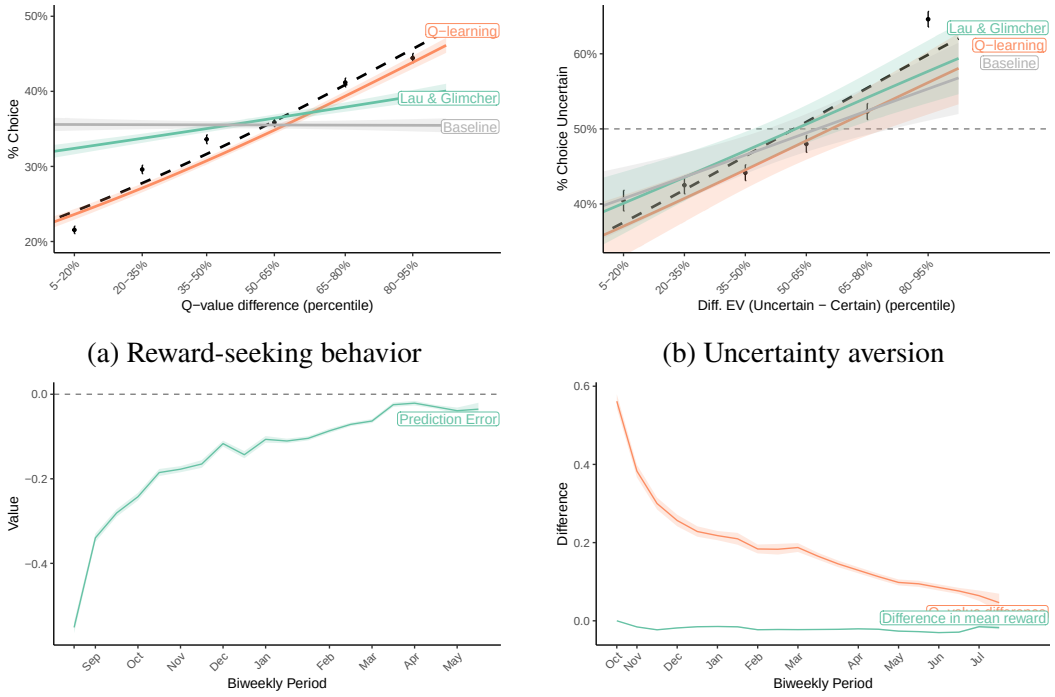
### Top Model Selection

These findings collectively suggest that teachers exhibit important reinforcement learning characteristics: they seek to maximize rewards and learn from experience over time. While the baseline model achieves better AIC scores in non-hierarchical estimation due to its parsimony, the hierarchical Q-learning model achieves superior out-of-sample prediction through LOOIC scores. Importantly, only the Q-learning models captures meaningful behavioral signatures that the baseline cannot explain. Thus, Q-learning provides the best fit among the theoretically informative models while also offering interpretable parameters related to learning and decision-making processes.

### Heterogeneity and Optimality

We analyzed teacher-specific parameters to capture individual differences in learning and the relationship between model parameters, classroom characteristics, and student performance metrics.

In the non-hierarchical estimation, we found that the cost parameter showed significant negative association with income level ($b = -0.260$, $p < .001$) and positive association with poverty level ($b = 0.218$, $p < .001$). The learning rate ($\alpha$) also showed a negative association with income level ($b = -0.119$, $p = .0156$), whereas the initial Q-value difference showed a positive association ($b = 0.232$, $p < .001$). Inverse temperature ($\tau$) and the discount factor ($\gamma$) did not show significant associations with classroom or school characteristics (see Table A.2).




(a) Reward-seeking behavior

(b) Uncertainty aversion

(c) Prediction Errors

(d) Q-value and reward differences

Figure 2.5: Behavioral signatures of reinforcement learning in teacher decision-making using non-hierarchical estimation.
The graphs compare three models (Q-learning, Logit, and Baseline) in their ability to capture various aspects of teachers' behavior. (a) Reward-seeking behavior: The x-axis represents the percentile of the difference in Q-values between action and inaction. The y-axis shows the proportion of times teachers chose to act. (b) Uncertainty aversion: The x-axis represents the percentile of the difference in expected value (EV) between uncertain and certain options, calculated from the cumulative means and standard deviations of rewards associated with each action. The y-axis shows the proportion of times teachers chose the uncertain option. (c) Prediction Errors: The plot shows the mean reward prediction errors across teachers over time. (d) Q-value and reward differences: The graph shows the difference in Q-values or mean rewards between action and inaction over time. In all plots, black points or dashed lines represent observed teacher behavior, while colored lines and shaded areas show model predictions with 95% confidence intervals.

To understand how these parameter differences relate to student outcomes, we examined their relationship with average weekly badges earned (indicating lesson completion). Tables 2.4 and 2.5 both show that the learning rate ($\alpha$) is positively associated with badges (non-hierarchical: $b = 0.072$, $p < .001$; hierarchical: $b = 0.063$, $p < .001$). However, the results diverge for inverse temperature ($\tau$) (non-hierarchical: $b = 0.064$, $p < .05$; hierarchical: $b = -0.090$, $p < .001$). Further, we


(a) Reward-seeking behavior

(b) Uncertainty aversion

(c) Prediction Errors

(d) Q-value and reward differences

Figure 2.6: Behavioral signatures of reinforcement learning in teacher decision-making using hierarchical estimation.
The graphs compare two hierarchical models (Q-learning and Baseline) in their ability to capture various aspects of teachers' behavior. (a) Reward-seeking behavior: The x-axis represents the percentile of the difference in Q-values between action and inaction. The y-axis shows the proportion of times teachers chose to act. (b) Uncertainty aversion: The x-axis represents the percentile of the difference in expected value (EV) between uncertain and certain options, calculated from the cumulative means and standard deviations of rewards associated with each action. The y-axis shows the proportion of times teachers chose the uncertain option. (c) Prediction Errors: The plot shows the mean reward prediction errors across teachers over time. (d) Q-value and reward differences: The graph shows the difference in Q-values or mean rewards between action and inaction over time. In all plots, black points or dashed lines represent observed teacher behavior, while colored lines and shaded areas show model predictions with 95% confidence intervals.

found differences in the strength of association with the discount factor ($\gamma$), which shows a strong negative relationship in the non-hierarchical model ($b = -0.082$, $p < .001$) but a weaker effect in the hierarchical case ($b = -0.044$, $p < .05$). Both models also show associations between badges and the starting Q-value difference (non-hierarchical: $b = -0.075$, $p =< .001$; hierarchical: $b = -0.089$, $p =< .001$). The cost parameter showed no consistent relationship with badges across models.

These associations remained robust when controlling for classroom characteristics (Models 2 and 3).

## 2.5 Discussion

Our study aimed to unravel the complex and adaptive nature of teacher behavior within the Zearn Math online platform. By leveraging reinforcement learning (RL) models, particularly the Q-learning approach, we sought to understand whether teachers demonstrate learning patterns that could be captured by these models.

### Characterizing Teacher Behavior

The Q-learning model demonstrated the best fit among hierarchical models, underscoring the importance of accounting for learning behaviors in this type of decision-making process. At a group level, the hierarchical Q-learning model's parameters may serve as a window into the diverse ways teachers generally navigate the digital learning environment. On the other hand, the non-hierarchical version of Q-learning seemed to overfit the data, highlighting a key limitation of our approach: the relatively sparse nature of our data (average 25.7 weekly observations per classroom) does not allow for robust fitting of non-hierarchical Q-learning models.

Our behavioral signature analysis also support reinforcement learning processes. The observed reward-seeking behavior, in which teachers show a preference for choices with higher relative expected values, aligned closely with the predictions of the Q-learning model. Another piece of evidence for learning was the consistent decrease in prediction error magnitudes over the academic year. This finding suggests that teachers were progressively improving their ability to anticipate the outcomes of their choices. However, when analyzing uncertainty aversion, we did not find significant differences between models, rendering this measure unsuccessful in falsifying any models.

### Impact of Estimated RL Parameters

Our analysis revealed a complex relationship between RL parameters and teacher or school characteristics. First, the positive association between the learning rate ($\alpha$) and student performance (badges) underscores the potential value of teacher adaptability. However, further research is needed to establish a causal relationship between educators who more readily update their expectations based on new information (including classroom performance) and their students' performance.

| | Dependent Variable: Badges | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| $\alpha$ | 0.064*** | 0.083*** | 0.072*** |
| | (0.015) | (0.015) | (0.015) |
| $\gamma$ | -0.072*** | -0.077*** | -0.082*** |
| | (0.018) | (0.017) | (0.018) |
| $\tau$ | 0.035* | 0.052 | 0.064* |
| | (0.017) | (0.027) | (0.027) |
| Cost | 0.004 | 0.021 | 0.027 |
| | (0.017) | (0.021) | (0.022) |
| Starting Q-value | -0.098*** | -0.073*** | -0.075*** |
| | (0.019) | (0.020) | (0.020) |
| No. of Weeks | | 0.024*** | 0.022*** |
| | | (0.002) | (0.003) |
| No. of Students | | 0.001 | 0.001 |
| | | (0.002) | (0.002) |
| No. of Classes | | -0.127*** | -0.126*** |
| | | (0.013) | (0.016) |
| Charter School | | | -0.086 |
| | | | (0.053) |
| Paid Zearn Account | | | 0.228*** |
| | | | (0.039) |
| Goodness of Fit (AIC) | | 0.169** | 0.204** |
| | | (0.064) | (0.066) |
| Constant | 1.558*** | 0.966*** | 1.357*** |
| | (0.013) | (0.140) | (0.210) |
| Control for Grade Level | | | Yes |
| Control for Poverty Level | | | Yes |
| Observations | 1,782 | 1,782 | 1,668 |
| $R^2$ | 0.067 | 0.154 | 0.210 |
| Adjusted $R^2$ | 0.064 | 0.149 | 0.202 |
| RSE (df) | 0.544 (1776) | 0.519 (1772) | 0.509 (1649) |
| F Statistic (df) | 25.514*** | 35.756*** | 24.404*** |
| | (5; 1776) | (9; 1772) | (18; 1649) |

*Note:* *p<0.05; **p<0.01; ***p<0.001

Table 2.4: Impact of Non-Hierarchical Q-learning Model Parameters on Average Weekly Badges Earned per Student.

Three linear regression models examine the associations between teacher-specific reinforcement learning (RL) parameters and student engagement, measured by average weekly badges earned per student. RL models were fit independently for each individual through maximum likelihood estimation. Model 1 includes only RL parameters. Model 2 controls for the goodness of fit of the Q-learning model for each teacher (AIC), the number of weeks, total students, and number of classes. Model 3 further incorporates controls for grade level, poverty level, charter school status, and whether the school has a paid Zearn account. Coefficients and standard errors (in parentheses) are provided for each parameter. RSE = Residual Standard Error.

| | Dependent Variable: Badges | | |
|---|---|---|---|
| | (1) | (2) | (3) |
| $\alpha$ | 0.062*** | 0.060*** | 0.063*** |
| | (0.015) | (0.015) | (0.015) |
| $\gamma$ | -0.032 | -0.045** | -0.044* |
| | (0.017) | (0.017) | (0.017) |
| $\tau$ | -0.039* | -0.098*** | -0.090*** |
| | (0.016) | (0.025) | (0.026) |
| Cost | -0.018 | -0.024 | -0.007 |
| | (0.021) | (0.021) | (0.021) |
| Starting Q-value | -0.149*** | -0.106*** | -0.089*** |
| | (0.021) | (0.024) | (0.025) |
| No. of Weeks | | 0.021*** | 0.018*** |
| | | (0.002) | (0.002) |
| No. of Students | | 0.002 | 0.001 |
| | | (0.002) | (0.002) |
| No. of Classes | | -0.125*** | -0.123*** |
| | | (0.013) | (0.016) |
| Charter School | | | -0.098 |
| | | | (0.052) |
| Paid Zearn Account | | | 0.220*** |
| | | | (0.039) |
| Goodness of Fit (LOOIC) | | 0.233 | 0.170 |
| | | (0.192) | (0.196) |
| Constant | 1.558*** | 1.446*** | 1.861*** |
| | (0.013) | (0.131) | (0.204) |
| Control for Grade Level | | | Yes |
| Control for Poverty Level | | | Yes |
| Observations | 1,782 | 1,782 | 1,668 |
| $R^2$ | 0.086 | 0.171 | 0.226 |
| Adjusted $R^2$ | 0.083 | 0.167 | 0.217 |
| RSE (df) | 0.538 (1776) | 0.513 (1772) | 0.504 (1649) |
| F Statistic (df) | 33.449*** | 40.660*** | 26.685*** |
| | (5; 1776) | (9; 1772) | (18; 1649) |

*Note:* *p<0.05; **p<0.01; ***p<0.001

Table 2.5: Impact of Hierarchical Q-learning Model Parameters on Average Weekly Badges Earned per Student.

Three linear regression models examine the associations between teacher-specific reinforcement learning (RL) parameters and student engagement, measured by average weekly badges earned per student. RL models were fit through hierarchical Bayes, where individual parameters are assumed to be drawn from population-level distributions. Model 1 includes only RL parameters. Model 2 adds controls for the goodness of fit of the Q-learning model for each teacher (LOOIC), number of weeks, total students, and number of classes. Model 3 further incorporates controls for grade level, poverty level, charter school status, and whether the school has a paid Zearn account. Coefficients and standard errors (in parentheses) are provided for each parameter. LOOIC = Leave-One-Out Information Criterion, RSE = Residual Standard Error.

Other findings showed less stability between hierarchical and non-hierarchical models and after controlling for school characteristics. The discrepancy in these results highlight the challenge of reliably estimating individual differences with limited observations. These results warrant further study to a) determine a direct relationship between RL parameters and teacher behavior, and b) understand why teachers with certain parameter values may have difference average student lesson completion.

**Influence of Teacher and School Background**

Our investigation also revealed associations between socioeconomic factors and teachers' interactions with Zearn Math. We found higher estimated cost parameters for teachers in schools with two different markers of low socioeconomic status (i.e., lower-income and high-poverty schools). This finding suggests a potential relationship between school socioeconomic status and teachers' perceived costs of implementing new teaching strategies. However, further research is needed to understand any mechanisms behind this association, such as whether resource constraints or differences in training adequacy affect the estimated cost parameter.

Further, we observed a negative correlation between school income levels and teachers' learning rates in our RL models. This relationship indicates that teachers in less affluent areas may demonstrate greater adaptability in adjusting their pedagogical approaches. While this finding is statistically significant, the relationship between resources and teaching adaptability is likely complex and influenced by many factors not captured in our model.

These findings point to a complex interplay between socioeconomic factors, school characteristics, and teachers' decision-making processes in the context of online learning platforms. Given the correlational nature of our study, future research could explore these relationships more deeply, potentially informing targeted interventions and support strategies. Such research might investigate how to ensure that all teachers, regardless of their school's socioeconomic status, have access to resources and training that could help them effectively adapt their teaching strategies and promote student success.

**Implications and Future Directions**

Our study represents an initial attempt to apply reinforcement learning models to complex field choice data. It paves the way for further understanding learning in real-world contexts, using frameworks derived from laboratory work. In general,

our Q-learning model highlights the dynamic, adaptive nature of teaching on the Zearn platform.

For educational practice, our results highlight the heterogeneity in optimal teaching strategies across educators, as revealed by our model parameters. This finding suggests that teachers may benefit from differentiated approaches to improving their teaching efforts.

From a policy perspective, the variations in model parameters across different school contexts highlight potential systemic educational disparities. While our study cannot establish causality, these findings raise important questions about the factors influencing teachers' decision-making and adaptability in various educational environments.

While our study provides valuable insights, it is not without limitations. A notable methodological challenged emerged in our model implementation, where data sparsity limited the non-hierarchical model estimates. Future work would benefit from longer observation periods and higher-frequency measurements for more robust estimation of individual differences.

Another significant limitation is the interpretability of our models. Eckstein et al. (2021) reviewed the interpretability and generalizability of reinforcement learning (RL) models in neuroscience and cognitive science, raising concerns on the widely adopted assumption that estimated RL parameters explain specific (neuro)cognitive functions across contexts. Methodological differences among RL studies yield a considerable variation in interpretation; for instance, learning rates have been linked to incremental updating, reward sensitivity, and approximate inference. This variability suggests caution in generalizing our findings across different contexts or populations. Future research should explore the applicability of our RL-based approach across different platforms, subject areas, and cultural contexts.

Our study also opens new avenues for exploring other RL models and methodologies to uncover deeper insights into the dynamics of educational technologies. These models, if applied correctly, could enhance our understanding of effective teaching strategies in digital contexts.

Additionally, while our model captures important aspects of teacher behavior, it does not account for all factors influencing the decision process. Future research could integrate a broader spectrum of variables, including teacher background and training, to provide a more comprehensive understanding of teaching and learning.

In conclusion, our study demonstrates the power of reinforcement learning models in uncovering the dynamic nature of teacher decision-making in a digital learning platform. By providing a new perspective on teacher behavior in online learning environments, we provide a foundation for future research and practice based on computational insights.

## 2.6 Materials and Methods

| Step | Methods | Software/Tools |
|---|---|---|
| Data Processing | Cleaning, Normalization | R (Team, 2024) (`tidyverse`, Wickham et al. (2019); `data.table`, Barrett et al. (2024)) |
| Dimensionality Reduction | Principal Component Analysis (PCA), Non-negative Matrix Factorization (NMF) | Python (`scikit-learn`, scikit-learn developers (n.d.)), R (`reticulate`, Ushey et al. (2024)) |
| Feature Selection, Analytical Methods | Q-learning Model Estimation | R (`cmdstanr` for Bayesian estimation, Gabry et al. (2024); `R.matlab`, Bengtsson (2022)), Matlab (`CBM` package for Laplace approximation, Piray et al. (2019)) |
| Model Evaluation | Heterogeneity analyses of model performance across teachers | R (`lmtest`, Zeileis and Hothorn (2002); `sandwich`, Zeileis (2006) and Zeileis et al. (2020)) |
| Visualization | Graphs and Tables | R (`ggplot2`, Wickham (2016); `gt`, Iannone et al. (2024); `stargazer`, Hlavac (2022)) |

Table 2.6: Analytical steps employed in the study.

### Data

Zearn provided administrative data for teachers and students, spanning across the 2020-2021 academic year. Teacher activity is time-stamped to the second and includes the time spent on the platform and specific actions taken. On the other hand, student data is aggregated at the classroom-week level due to data privacy considerations. As such, we aggregated the teacher data to the classroom-week unit of analysis. This level of granularity still enabled us to capture the temporal dynamics of teacher-student interactions and their subsequent influence on student achievement.

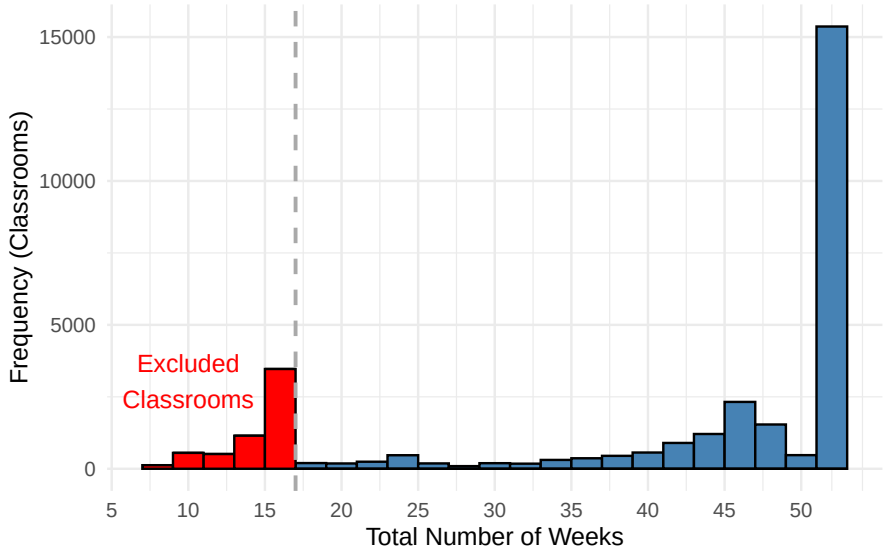
| | Mean | Median | SD | Min | Max |
|---|---|---|---|---|---|
| Teachers | 12.09 | 9 | 11.86 | 1 | 72 |
| Students | 268.69 | 207 | 279.57 | 1 | 3,289 |
| Weeks | 24.10 | 27 | 12.54 | 1 | 51 |

Table 2.7: Summary statistics by school.
The table presents the mean, median, standard deviation (SD), minimum, and maximum values for the number of teachers, total students, and average weeks of active engagement (across all classrooms within a school).

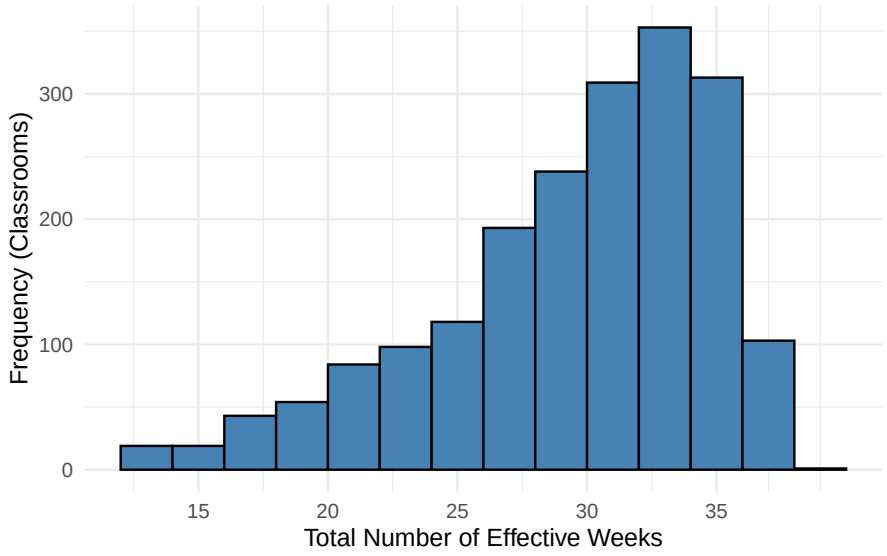
The dataset includes 31,046 classrooms and 19,689 educators, with an average of 17.6 students per classroom. Classrooms and teachers are also linked to a school, and Table 2.7 provides a summary of the number of students, teachers, and weeks per school (see also Figures A.4a and A.4b for the distributions of school median poverty and income levels).

**Preprocessing and Exclusion criteria**

We focused our analysis on the teachers who most likely take advantage of a wide range of resources on the platform. Thus, we selected teachers who consistently use the platform and work in traditional school settings. First, we selected virtual classrooms with at least five active students weekly, filtering out parents or tutors who may use Zearn outside the classroom setting. We also removed teachers with more than four classrooms and those who logged in for less than 16 weeks (Figure 2.7a reveals that a non-negligible number of classrooms has less than three to four months of data). Finally, we excluded classrooms in the 6th to 8th grades, as they represent only a small proportion of the data. Table 2.8 summarizes the refined dataset, providing a snapshot of the key variables of interest. Their means and standard deviations (SD) are computed for each grade level and overall (across all grades).




(a) Before data cleaning




(b) After data cleaning

Figure 2.7: Histograms of number of weeks of data per classroom.

Panel (a) shows the raw distribution of the number of weeks per classroom. Most classrooms include a full year (52 weeks) of data. A smaller but significant subset of classrooms has less than 18 weeks of data. The dashed line represents the exclusion threshold. Some classrooms consistently use the platform throughout the academic year, while others show sporadic engagement, possibly reflecting trial periods or intermittent usage. Panel (b) displays the distribution of the number of weeks per classroom after data cleaning.

| | Minutes | Badges | Tower Alerts | Teacher Minutes |
|---|---|---|---|---|
| Overall, N = 135,784 | 81 (56) | 2.43 (1.91) | 0.47 (0.81) | 84 (149) |
| Kindergarten, N = 7,486 | 41 (34) | 4.44 (3.59) | 0.05 (0.35) | 21 (46) |
| 1st, N = 21,126 | 80 (51) | 2.50 (1.64) | 0.40 (0.45) | 65 (107) |
| 2nd, N = 26,424 | 84 (55) | 2.46 (1.67) | 0.36 (0.59) | 70 (122) |
| 3rd, N = 26,316 | 84 (55) | 2.42 (1.71) | 0.44 (0.60) | 100 (175) |
| 4th, N = 27,048 | 84 (59) | 2.13 (1.74) | 0.56 (0.81) | 100 (169) |
| 5th, N = 27,384 | 84 (59) | 2.10 (1.66) | 0.71 (1.28) | 101 (165) |

Table 2.8: Summary statistics by grade level.
Classroom engagement metrics by grade level. The table presents the means and standard deviations (in parentheses) of the following averages: minutes spent on the platform per student per week, badges earned per student per week (indicating lesson completion), Tower Alerts per lesson completion (indicating student struggle), and minutes spent on the platform per teacher per week.

## Operationalizing Actions, Rewards, and States

### Teacher Actions

Teacher actions encompass a broad spectrum, from platform log-ins to resource downloads and specific instructional activities. Table A.1 provides a list of the actions available in the data.

### Rewards (Student Actions)

In reinforcement learning (RL) models, reward variables capture the dynamics of the environment in which learning and decision-making occur. The Zearn platform provides a rich set of student activity and performance data that can be used to define these variables, offering a quantifiable snapshot of classroom engagement and learning challenges. Reward variables quantify the desirability of the outcomes resulting from the agent's actions, serving as feedback signals that guide the learning process.

In this study, we take an agnostic approach, allowing the following student variables to be treated as reward variables depending on our RL model specification:

1. "Active Students": This variable represents the number of students actively logging in to complete digital lessons within a given week (Zearn, 2022). A high number of active students could be considered a positive outcome, indicating successful student engagement.

2. "Student Logins": This variable tallies the frequency of students entering the platform, potentially serving as an engagement metric (Zearn, 2024a).A high frequency of student logins could be viewed as a positive outcome, reflecting consistent student engagement.

3. "Badges (on grade)" and "Badges": These metrics reflect the number of new lessons completed weekly at students' grade level and in general (Zearn, 2024e). Accumulating badges can serve as a positive signal, indicating students' mastery of the curriculum.

4. "Minutes per active student": This variable measures students' time on the platform, potentially correlating with their focus and learning progress (Zearn, 2022). Achieving or exceeding the expected minutes can be considered a positive outcome, while falling short may be viewed as a negative outcome.

5. "Tower Alerts": These alerts signal instances when students repeatedly encounter difficulties within the same lesson (Zearn, 2024f). Tower Alerts could be viewed as a negative signal, indicating that the current teaching strategies may not be effective in addressing student difficulties.

**Dimensionality Reduction**

First, we standardized the dataset by z-scoring the variables of interest at the school level (using school-wide means and standard deviations). We performed NMF and evaluated the data's reconstruction accuracy and cluster separation using, respectively, the sum of squared residuals (a measure of the difference between the original data and the reconstructed data) and silhouette scores—a measure of how similar an object is to its cluster compared to other clusters (Rousseeuw, 1987).

We calculate the silhouette score with the formula $(b - a)/\max(a, b)$, where $a$ is the average distance within a cluster and $b$ is the average distance to the nearest neighboring cluster. This score ranges from -1 to 1, with higher values indicating a data point is well-matched to its cluster and poorly matched to neighboring clusters.

**Nonnegative Matrix Factorization (NMF) Methodology**

Let the original matrix ($\mathbf{X}$) be a detailed description of all the teachers' (or students') behaviors. Each row in the matrix represents a unique teacher-week (or classroom-week), and each column represents a specific behavior or action. The entry in

a specific row and column corresponds to the frequency of that behavior for that particular teacher-week (or classroom-week). We then estimate $\mathbf{X} \simeq \mathbf{WH}$, such that we minimize the following:

$$\|\mathbf{X} - \mathbf{WH}\|, \mathbf{W} \geq 0, \mathbf{H} \geq 0.$$

We used two different loss functions (Frobenius norm and Kullback-Leibler divergence) and two different initialization methods (nonnegative double singular value decomposition (NNDSVD) and NNDSVD with zeros filled with the average of the input matrix (NNDSVDA)). The resulting matrices are:

1. Basis Matrix ($\mathbf{W}$): This matrix represents underlying behavior patterns. Each column is a meta-behavior or a group of behaviors occurring together.
2. Mixture Matrix ($\mathbf{H}$): This matrix shows the extent to which each meta-behavior is present in each teacher-week (or classroom-week). Each entry in this matrix represents the contribution of a meta-behavior to a particular behavior present in the data.

These matrices can reveal underlying patterns of behaviors (from the basis matrix) and how these patterns are mixed and matched in different teachers (from the mixture matrix). It allows us to assess the method's performance under varying configurations, with the sum of squared residuals and silhouette scores for comparison.

**Model Performance and Feature Selection**

To identify the most appropriate action and reward variables, we employed a multiple model types:

1. Baseline Model:

$$\text{Action}_t = \beta_0 + \epsilon_t$$

where $\beta_0$ is the intercept and $\epsilon_t$ is the error term.

2. Logistic Regression Model: Inspired by dynamic analysis (Lau & Glimcher, 2005), incorporating lagged variables to capture temporal dynamics:

$$\text{logit}(P(\text{Action}_t = 1)) = \beta_0 + \sum_{i=1}^{L} (\beta_i R_{t-i} + \gamma_i \text{Action}_{t-i})$$

where $L$ is the number of lags, $R_t$ is the reward at time $t$, and $\beta_i$, and $\gamma_i$ are coefficients.

3. Q-learning Model: A reinforcement learning model capturing adaptive decision-making processes, as explained in the theory section.

4. Simplified Q-learning Model: A version of the Q-learning model with no cost parameter and a starting Q-value of 0.

We applied reward structures extracted from classroom data via Non-negative Matrix Factorization (NMF) with the Frobenius Non-negative Double Singular Value Decomposition (NNDSVD). Actions were derived similarly from teacher data. We selected one teacher component as the action and one student component as the reward, resulting in 16 possible configurations (4 possible actions and 4 possible rewards).

**Model Estimation**

In our analysis, we adopt two estimation approaches. First, we use individual maximum likelihood estimation, as described by Piray et al. (2019), to assess the fitness of our RL models and estimate individual parameters across teachers.

Secondly, we implement hierarchical Bayesian inference in Stan using Hamiltonian Monte Carlo sampling. Within this framework, we assume that for any given model $m$ and teacher $n$, the individual parameters ($h_{m,n}$) are normally distributed across the population with $h_{m,n} \sim N(\mu_m, \sigma_m)$, where $\mu_m$ and $\sigma_m$ represent the vectors of means and standard deviation of the distribution over $h_{m,n}$, respectively. We use half-Cauchy priors for population standard deviations and non-centered parameterization (see Betancourt and Girolami (2015)).

For both hierarchical and non-hierarchical models, we transform the initial estimates to generate constrained model parameters (e.g., the learning rate and discount factor in the Q-learning model). For parameters within a $(0, 1)$ interval, we use the inverse logit function transform, $\text{Logit}^{-1}(x) = 1/(1+e^{-x})$, and for intrinsically non-negative parameters, we use an exponential transformation. Consequently, we estimate the following unconstrained parameters:

1. Q-learning:
    - Learning Rate: $\text{Logit}(\alpha) = \log(\frac{\alpha}{1-\alpha})$
    - Discount Rate: $\text{Logit}(\gamma) = \log(\frac{\gamma}{1-\gamma})$
    - Inverse Temperature: $\log(\tau)$
    - Cost: $\log(\text{cost})$
    - Initial Q-value: $\Delta Q_{t=0}$
2. Logistic and Baseline Regression Models:
    - Parameters: $\beta$

**Top Model Selection**

Similar to Charpentier et al. (2024), we determined the best-fit model from our set of non-hierarchical candidates by considering each model's Akaike Information Criterion (AIC). We computed the AIC using the maximum likelihood parameter estimates:

$$\text{AIC} = \sum_{i=1}^{N} \left(2p - 2\ln(\hat{L}_i)\right)$$

where $N$ is the number of teachers, $p$ is the number of model parameters, and $\hat{L}_i$ is the maximum likelihood estimate for teacher $i$ given their individual parameter estimates.

For the hierarchical Bayesian estimation, we used the Leave-One-Out Information Criterion (LOOIC) with mixture importance sampling (MixIS), following Silva and Zanella (2024). This approach estimates out-of-sample prediction accuracy by sampling from a mixture of leave-one-out posteriors:

$$q_{mix}(\theta) = \frac{\sum_{i=1}^{n} p(y_{-i}|\theta)p(\theta)}{\sum_{i=1}^{n} p(y_{-i})} \propto p(\theta|y) \cdot \left(\sum_{i=1}^{n} p(y_i|\theta)^{-1}\right)$$

where $p(y_{-i}|\theta)$ is the likelihood excluding observation $i$. The LOOIC is then computed from the resulting estimates of LOO predictives, that is, $\text{LOOIC} = -2\sum_{i=1}^{n} p(y_i|y_{-i})$.

**Behavioral Signatures**

We examined three key behavioral signatures, similar to Charpentier et al. (2024) and Cockburn et al. (2022):

1. Reward Seeking: We calculated the probability of choosing an action as a function of its Q-value difference (percentile rank of Q-values). For each model (Q-learning, Lau & Glimcher, and Baseline), we fitted a generalized linear model (GLM) with a binomial family. The model took the form:

$$\text{logit}(P(\text{Action})) = \beta_0 + \beta_1 Q_{\text{percentile}}$$

where $Q_{\text{percentile}}$ is the percentile rank of Q-values. We extracted the slope ($\beta_1$) and its 95% confidence interval using clustered standard errors at the classroom level to account for within-classroom correlations.

2. Uncertainty Aversion: We examined the probability of choosing the uncertain option as a function of the difference in expected value (EV) between uncertain and certain options. Uncertainty was defined based on the cumulative standard deviation of rewards associated with each action. We fitted a GLM with a binomial family:

$$\text{logit}(P(\text{Uncertain Choice})) = \beta_0 + \beta_1 \Delta EV_{\text{percentile}}$$

where $\Delta EV_{\text{percentile}}$ is the percentile rank of the difference in expected value between uncertain and certain options. We calculated the indifference point (where $P(\text{Uncertain Choice}) = 0.5$) and its 95% confidence interval for each model.

3. Learning Dynamics: We evaluated learning over time by examining two aspects. First, we computed mean prediction errors across teachers as:

$$PE_t = \gamma R_t - Q_{t-1}(a)$$

where $\gamma$ is the discount factor, $R_t$ is the reward at time $t$, and $Q_{t-1}(a)$ is the Q-value of the chosen action at the previous time step. We then evaluated the evolution of Q-value differences and the difference in mean rewards between action and inaction over time. The reward difference was calculated using a rolling mean with a window size of 4 weeks.

All statistical analyses were performed using the `lmtest` package (Zeileis & Hothorn, 2002) for clustered standard errors and the sandwich package (Zeileis, 2004) for robust covariance matrix estimation.

**Heterogeneity Analysis**

After selecting the top-performing model (Q-learning), we explored heterogeneity through:

1. Parameter-Classroom Characteristic Associations: For each Q-learning parameter (learning rate $\alpha$, discount factor $\gamma$, inverse temperature $\tau$, initial Q-value, and cost), we fitted separate models with the following specifications:

    a) For ordinal outcomes (income and poverty levels), we used ordered logistic regression:

    $$\text{logit}(P(Y \leq j)) = \theta_j - (\beta_1 X_1 + ... + \beta_p X_p)$$

    b) For count outcomes (total students and number of classes), we employed Poisson regression:

    $$\log(E(Y)) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

    c) For binary outcomes (paid account status), we used logistic regression:

    $$\text{logit}(P(Y = 1)) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p$$

    where $X_1, ..., X_p$ represent the standardized Q-learning parameters.

2. Parameter-Student Outcome Associations: We examined how model parameters relate to student outcomes using multiple linear regression models. Two key outcomes were analyzed:

    a) Average weekly badges earned per student (indicating lesson completion)
    b) Average weekly tower alerts per tower completion (indicating student struggles)

    The regression model took the form:

    $$Y = \beta_0 + \beta_1 \alpha + \beta_2 \gamma + \beta_3 \tau + \beta_4 Q_0 + \beta_5 \text{cost} + C_i \Gamma + \epsilon$$

where $Y$ is the outcome variable, $\alpha$, $\gamma$, $\tau$, $Q_0$, and cost are the Q-learning parameters, and $C_i$ is a matrix of control variables including AIC, number of weeks, total students, number of classes, grade level, poverty level, charter school status, and paid account status.

## References

Atkinson, R. C. (1972). Optimizing the learning of a second-language vocabulary [Place: US Publisher: American Psychological Association]. *Journal of Experimental Psychology*, *96*(1), 124–129. https://doi.org/10.1037/h0033475

Barrett, T., Dowle, M., Srinivasan, A., Gorecki, J., Chirico, M., & Hocking, T. (2024). *Data.table: Extension of 'data.frame'*. manual. https://r-datatable.com

Bellman, R. (2015). *Adaptive control processes: A guided tour* [Citation Key: bellman2015adaptive]. Princeton University Press. https://books.google.com/books?id=rLnEnQEACAAJ

Bengtsson, H. (2022). *R.matlab: Read and write MAT files and call MATLAB from within R*. manual. https://github.com/HenrikBengtsson/R.matlab

Betancourt, M., & Girolami, M. (2015). Hamiltonian Monte Carlo for Hierarchical Models. In S. K. Upadhyay, U. Singh, D. Dey, & A. Loganathan (Eds.), *Current trends in Bayesian methodology with applications* (p. 24). CRC Pres.
OCLC: 910237218.

Brown, V. M., Zhu, L., Solway, A., Wang, J. M., McCurry, K. L., King-Casas, B., & Chiu, P. H. (2021). Reinforcement learning disruptions in individuals with depression and sensitivity to symptom change following cognitive behavioral therapy [49 citations (Crossref) [2024-05-30] 43 citations (Semantic Scholar/DOI) [2024-04-15]]. *JAMA Psychiatry*, *78*(10), 1113. https://doi.org/10.1001/jamapsychiatry.2021.1844

Charpentier, C. J., Wu, Q., Min, S., Ding, W., Cockburn, J., & O'Doherty, J. P. (2024). Heterogeneity in strategy use during arbitration between experiential and observational learning. *Nature Communications*, *15*(1), 4436. https://doi.org/10.1038/s41467-024-48548-y

Cockburn, J., Man, V., Cunningham, W. A., & O'Doherty, J. P. (2022). Novelty and uncertainty regulate the balance between exploration and exploitation through distinct mechanisms in the human brain. *Neuron*, *110*(16), 2691–2702.e8. https://doi.org/10.1016/j.neuron.2022.05.025

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876–879. https://doi.org/10.1038/nature04766

Dennison, J. B., Sazhin, D., & Smith, D. V. (2022). Decision neuroscience and neuroeconomics: Recent progress and ongoing challenges [15 citations (Crossref) [2024-06-07]]. *WIREs Cognitive Science*, *13*(3), e1589. https://doi.org/10.1002/wcs.1589

Doroudi, S., Aleven, V., & Brunskill, E. (2019). Where's the reward?: A review of reinforcement learning for instructional sequencing. *International Journal*

*of Artificial Intelligence in Education*, *29*(4), 568–620. https://doi.org/10.1007/s40593-019-00187-x

Eckstein, M. K., Wilbrecht, L., & Collins, A. G. (2021). What do reinforcement learning models measure? interpreting model parameters in cognition and neuroscience [55 citations (Crossref) [2024-05-30] 49 citations (Semantic Scholar/DOI) [2024-04-15]]. *Current Opinion in Behavioral Sciences*, *41*, 128–137. https://doi.org/10.1016/j.cobeha.2021.06.004

Gabry, J., Češnovar, R., Johnson, A., & Bronder, S. (2024). *Cmdstanr: R Interface to 'CmdStan'*. manual. https://mc-stan.org/cmdstanr/

Hagger, M. S., Hamilton, K., Phipps, D. J., Protogerou, C., Zhang, C.-Q., Girelli, L., Mallia, L., & Lucidi, F. (2023). Effects of habit and intention on behavior: Meta-analysis and test of key moderators. *Motivation Science*, *9*(2), 73–94. https://doi.org/10.1037/mot0000294

Hlavac, M. (2022). *Stargazer: Well-formatted regression and summary statistics tables*. manual. Bratislava, Slovakia, Social Policy Institute. https://CRAN.R-project.org/package=stargazer

Howard, R. A. (1960). *Dynamic programming and Markov processes*. Technology Press of Massachusetts Institute of Technology.

Iannone, R., Cheng, J., Schloerke, B., Hughes, E., Lauer, A., & Seo, J. (2024). *Gt: Easily create presentation-ready display tables*. manual. https://gt.rstudio.com

Jumaat, N. F., & Tasir, Z. (2014). 2014 international conference on teaching and learning in computing and engineering (latice), 74–77. https://doi.org/10.1109/LaTiCE.2014.22

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey [4495 citations (Crossref) [2024-05-30] 8379 citations (Semantic Scholar/DOI) [2024-04-15] QID: Q61773137]. *Journal of Artificial Intelligence Research*, *4*, 237–285. https://doi.org/10.1613/jair.301

Knudsen, J., Lundh, P., Hsiao, M., & Saucedo, D. (2020, December). *Zearn math curriculum study professional development final report* (tech. rep.).

Lau, B., & Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the Experimental Analysis of Behavior*, *84*(3), 555–579. https://doi.org/10.1901/jeab.2005.110-04

Memarian, B., & Doleck, T. (2024). A scoping review of reinforcement learning in education. *Computers and Education Open*, *6*, 100175. https://doi.org/10.1016/j.caeo.2024.100175

Morrison, J., Wolf, B., Ross, S., Risman, K., & McLemore, C. (2019, April). *Efficacy study of Zearn math in a large urban school district*. Center for Research and Reform in Education. http://jhir.library.jhu.edu/handle/1774.2/62395

Niv, Y., Hitchcock, P., Berwian, I. M., & Schoen, G. (2022). Toward precision cognitive behavioral therapy via reinforcement learning theory. In L. M. Williams & L. M. Hack (Eds.). American Psychiatric Association Publishing.

Palminteri, S., Wyart, V., & Koechlin, E. (2017). The Importance of Falsification in Computational Cognitive Modeling. *Trends in Cognitive Sciences*, *21*(6), 425–433. https://doi.org/10.1016/j.tics.2017.03.011

Park, H. W., Grover, I., Spaulding, S., Gomez, L., & Breazeal, C. (2019). A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education [76 citations (Crossref) [2024-05-30] 93 citations (Semantic Scholar/DOI) [2024-04-15]]. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 687–694. https://doi.org/10.1609/aaai.v33i01.3301687

Piray, P., Dezfouli, A., Heskes, T., Frank, M. J., & Daw, N. D. (2019). Hierarchical bayesian inference for concurrent model fitting and comparison for group studies (H. Berry, Ed.) [60 citations (Crossref) [2024-03-29]]. *PLOS Computational Biology*, *15*(6), e1007043. https://doi.org/10.1371/journal.pcbi.1007043

Reiser, B. J., & Tabak, I. (2014, September 23). Scaffolding [DOI: 10.1017/CBO9781139519526.005]. In R. K. Sawyer (Ed.). Cambridge University Press. https://www.cambridge.org/core/product/identifier/9781139519526%23c03325-3-1/type/book_part

Rescorla, R., & Wagner, A. R. (1972). A theory of pavlovian conditioning : Variations in the effectiveness of reinforcement and nonreinforcement [Citation Key: Rescorla1972ATO]. https://api.semanticscholar.org/CorpusID:51139715

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis [11569 citations (Crossref) [2024-03-29]]. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

scikit-learn developers. (n.d.). *2.3. clustering* [Scikit-learn]. Retrieved April 23, 2024, from https://scikit-learn.org/stable/modules/clustering.html

Silva, L. A., & Zanella, G. (2024). Robust Leave-One-Out Cross-Validation for High-Dimensional Bayesian Models. *Journal of the American Statistical Association*, *119*(547), 2369–2381. https://doi.org/10.1080/01621459.2023.2257893

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (Second edition). The MIT Press.

Team, R. C. (2024). *R: A language and environment for statistical computing* (tech. rep.). Vienna, Austria. https://www.R-project.org/

Thorndike, E. L. (1931). *Human learning*. [DOI: 10.1037/11243-000]. The Century Co. https://doi.org/10.1037/11243-000

Ushey, K., Allaire, J. J., & Tang, Y. (2024). *Reticulate: Interface to 'Python'*. manual. https://rstudio.github.io/reticulate/

Venkatesh, V., Davis, F. D., & Zhu, Y. (2023). Competing roles of intention and habit in predicting behavior: A comprehensive literature review, synthesis, and longitudinal field study [7 citations (Crossref) [2024-06-07]]. *International Journal of Information Management*, *71*, 102644. https://doi.org/10.1016/j.ijinfomgt.2023.102644

Verplanken, B., & Orbell, S. (2022). Attitudes, habits, and behavior change [88 citations (Crossref) [2024-06-07]]. *Annual Review of Psychology*, *73*(1), 327–352. https://doi.org/10.1146/annurev-psych-020821-011744

Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3-4), 279–292. https://doi.org/10.1007/BF00992698

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis* (Second edition). Springer.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., . . . Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wyatt, L. E., Hewan, P. A., Hogeveen, J., Spreng, R. N., & Turner, G. R. (2024). Exploration versus exploitation decisions in the human brain: A systematic review of functional neuroimaging and neuropsychological studies. *Neuropsychologia*, *192*, 108740. https://doi.org/10.1016/j.neuropsychologia.2023.108740

Zearn. (2022). Independent digital lessons: "look fors" guide (g1-8).

Zearn. (2024a). *Download student logins* [Zearn help center]. Retrieved March 29, 2024, from https://help.zearn.org/hc/en-us/articles/5700246424343-Download-student-logins

Zearn. (2024b). *Efficacy research* [Zearn math]. Retrieved June 29, 2023, from https://about.zearn.org/research

Zearn. (2024c). *How to get started* [Zearn math]. Retrieved July 16, 2024, from https://about.zearn.org/getting-started

Zearn. (2024d). *How zearn works* [Math curriculum & digital lessons]. Retrieved July 16, 2024, from https://about.zearn.org/how-zearn-math-works

Zearn. (2024e). *Independent digital lesson components* [Zearn help center]. Retrieved March 29, 2024, from https://help.zearn.org/hc/en-us/articles/236238988-Independent-Digital-Lesson-components

Zearn. (2024f). *Independent practice: Tower of power* [Zearn help center]. Retrieved January 31, 2024, from https://help.zearn.org/hc/en-us/articles/360052426593-Independent-Practice-Tower-of-Power

Zearn. (2024g). *Lesson materials* [Zearn help center]. Retrieved March 29, 2024, from https://help.zearn.org/hc/en-us/articles/4403432269719-Lesson-Materials

Zearn. (2024h). *Locked student lessons* [Zearn help center]. Retrieved July 16, 2024, from https://help.zearn.org/hc/en-us/articles/115008107568-Locked-student-lessons

Zearn. (2024i). *Tower alerts report* [Zearn help center]. Retrieved March 29, 2024, from https://help.zearn.org/hc/en-us/articles/236239748-Tower-Alerts-Report

Zeileis, A. (2006). Object-Oriented Computation of Sandwich Estimators. *Journal of Statistical Software*, *16*(9). https://doi.org/10.18637/jss.v016.i09

Zeileis, A., & Hothorn, T. (2002). Diagnostic checking in regression relationships. *R News*, *2*(3), 7–10. https://CRAN.R-project.org/doc/Rnews/

Zeileis, A., Köll, S., & Graham, N. (2020). Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in *R*. *Journal of Statistical Software*, *95*(1). https://doi.org/10.18637/jss.v095.i01

Zhang, L., Lengersdorff, L., Mikus, N., Gläscher, J., & Lamm, C. (2020). Using reinforcement learning models in social neuroscience: Frameworks, pitfalls and suggestions of best practices [78 citations (Crossref) [2024-06-07]]. *Social Cognitive and Affective Neuroscience*, *15*(6), 695–707. https://doi.org/10.1093/scan/nsaa089

*Chapter 3*

# DISCOVERING DATA-DRIVEN NUDGES TO HELP STUDENTS LEARN MORE MATH

## 3.1 Abstract

The COVID-19 pandemic has exacerbated learning deficits, particularly in mathematics (Di Pietro, 2023). Adaptive online learning platforms offer a promising strategy to mitigate these adverse effects (Meeter, 2021), but designing effective interventions to improve student outcomes via such platforms remains a challenge. Here, we present a two-phase study that uses granular teacher and student engagement data from a large online Math platform in the United States. First, using unsupervised learning techniques (Independent Component Analysis, ICA; and a modified Predicting Context Sensitivity (PCS) approach, drawn from Buyalskaya et al. (2023), we identified critical teacher behaviors associated with improved student performance, namely empathy-focused engagement and strategic weekly login patterns. Building on these insights and in consultation with instructors, we developed two behaviorally-informed interventions: an empathy intervention encouraging teachers to view math problems from the student's perspective, and a habit-building intervention emphasizing the importance of Friday logins for reflective analysis and proactive planning. In a large-scale randomized controlled trial (N = 140,461 teachers across 22,281 schools), we demonstrated that the empathy intervention significantly increased student lesson completion by 4.09%. The habit condition also showed promise, increasing lesson completion by 1.81%, although this effect was not significantly different from the active control. Our approach demonstrates the potential of data-driven behavioral interventions to inform the development of more effective digital learning interventions, ultimately improving math education outcomes for students across diverse educational settings. Furthermore, we provide a generalizable framework for designing targeted interventions in various contexts where granular behavioral data are available.

## 3.2 Introduction

The decline in math performance among American students has been a critical issue exacerbated by the COVID-19 pandemic, with some reporting an alarming half-year lag in math achievement among U.S. public school students in grades 3-8 (Fahle et al.,

2023). Learning losses and disparities, particularly among younger students, have been significant due to reduced instructional time and remote learning challenges (Di Pietro, 2023; Zierer, 2021). Moreover, Ewing and Green (2021) note that repeated school closures compounded these issues, emphasizing that the pandemic is not the only problem affecting math performance.

Emphasizing quality math instruction, particularly in pedagogical strategies, guidance, and teacher-student relations, is crucial for improving math performance, especially for students who have lost interest in mathematics or lack a sufficient foundation (Battey et al., 2016; Wang et al., 2023). Furthermore, teachers' deep mathematical knowledge is also important, enabling them to understand students' challenges better and provide adequate support (Battey et al., 2016; Hill et al., 2008). This combination of empathy and subject expertise highlights the importance of specialized training focusing on pedagogical skills and content knowledge.

The pandemic also accelerated the adoption of digital platforms for math education. Meeter (2021) found that adaptive practice software effectively mitigated the adverse effects of school closures. This result is consistent with the qualitative evidence from Alabdulaziz (2021), who reported that teachers found digital platforms beneficial in addressing the challenges of remote learning during the pandemic. Recent meta-analyses demonstrate that integrating digital tools and blended learning approaches improves student outcomes significantly (Ran et al., 2020, 2021; Sadaf et al., 2021). Leveraging these tools and insights gained during the pandemic should help address longstanding educational challenges (Ewing & Green, 2021). In particular, integrating technology, pedagogy, and content knowledge is essential, and professional development programs are most effective when they focus on using technology to foster a more engaging and effective learning environment (Blanchard et al., 2016; Young, 2016).

The increased use of digital platforms has provided more data to support the importance of student engagement in online learning. Blending online and in-person teaching methods can effectively enhance engagement and understanding, depending on the implementation (Chiang et al., 2016; Q. Li & Ma, 2010; S. Li & Wang, 2022; Mawardi et al., 2023; Sadaf et al., 2021; Yu et al., 2023). Teachers play a crucial role in helping students develop meta-cognitive skills to foster student engagement (Haleva et al., 2020). Furthermore, teachers' beliefs and self-efficacy toward technology integration influence their willingness to adopt innovative teaching practices (Ertmer et al., 2012; Liljedahl & Oesterle, 2020). This relationship

between student engagement and teacher attitudes suggests the importance of creating solutions that can be implemented across diverse educational settings. Our current study aims to address this need by developing and evaluating cost-effective, high-quality interventions designed to improve student outcomes through enhanced teacher engagement with the Zearn platform.

In this study, we partner with Zearn to address these topics. This math education platform reaches approximately 25% of elementary schools and over a million middle-school students in the United States. Zearn's approach integrates interactive digital lessons with hands-on teaching, aligning with the Common Core State Standards and providing a comprehensive educational experience ("Zearn Math: Top-Rated Math Learning Platform", 2023).

Our study leverages this rich resource to offer an innovative approach to educational interventions using behavioral science principles. Focusing on quality of teaching, we align with Hanushek (2020), who maintains that the efficacy of resource utilization supersedes quantity. We also follow current trends in providing cost-effective, easy-to-implement interventions (i.e., *nudges*) that offer engagement incentives for both teachers and students (Koch et al., 2015; Lavecchia et al., 2016; Lynch et al., 2019).

Our two-step approach initially employs unsupervised learning techniques to analyze behavioral patterns in teacher activity on Zearn, aligning with the data mining value in educational research (Al-Shabandar et al., 2018; Hershcovits et al., 2020; Qiu et al., 2022; Salazar et al., 2007; Shin & Shim, 2020). Subsequently, we aim to establish the causality of our interventions through a large-scale field experiment guided by recommendations from Greene (2022) for holistic, transparent, and reproducible research. Our unique integration of behavioral science with digital education seeks to provide impactful insights into math education and offer a blueprint for similar studies in other fields.

**Hypothesis**

We hypothesized that interventions designed based on insights from our data analysis would lead to significant improvements in student learning outcomes.

## 3.3 Results

### Study 1: Data-Driven Nudge Design

Our study used a comprehensive dataset from Zearn's educational platform, encompassing the 2019-2020 academic year and spanning multiple schools in Louisiana. Zearn's content is designed to promote intuitive understanding by progressing from concrete to pictorial to abstract examples. The platform offers a personalized experience, providing targeted remediation when a student encounters difficulties. Teachers can monitor individual student progress by tracking activities such as *Badges* and *Tower of Power*. Badges track student lesson completion, and the Tower of Power is an assessment that presents students with challenging problems at the end of each lesson. When students struggle or fail to answer correctly during a Tower of Power, the platform provides *Boosts* (i.e., hints or further explanations) and notifies teachers with a *Tower Alert*. These features aim to motivate students and provide valuable information for educators to support learning.

Other important variables from the dataset included teacher logins, file downloads, and specific interactions with educational content. Additionally, we accessed student data at the classroom-week level, encompassing metrics such as lesson completion (i.e., Badges) and instances of learning difficulties (i.e., Tower Alerts). This granularity allowed for an in-depth analysis of both teacher behaviors and student performance.

The data revealed diverse engagement patterns. Teachers logged in multiple times per week, exhibiting variations in the frequency and duration of their interactions. The student data, aggregated at the classroom-week level, showed a wide range of performance levels across various classrooms and schools. The standard deviations of key variables underscored this diversity, as detailed in the summary statistics of Table 3.1. This rich combination of teacher behaviors and student performance metrics, carefully matched with a weekly frequency for each classroom, allowed for an in-depth analysis while upholding privacy standards. Note that the analyses in Study 1 underwent post hoc modifications to rectify some subsequently identified errors. These changes, while adjusting the coefficients slightly, did not significantly alter the overall results or patterns observed (refer to the SI for the original analyses).

Table 3.1: Key educational metrics

| | N | Aggregated Statistics per Teacher Mean Standard | | | | |
|---|---|---|---|---|---|---|
| | | Mean | SD | 1st Quartile | Median | 3rd Quartile |
| No. of teachers | 2,506 | - | - | - | - | - |
| No. of classes | 4,228 | 1.69 | 0.97 | 1.0 | 1.0 | 2.0 |
| No. of badges | 5,230,607 | 2,087.23 | 1,627.26 | 1,035.2 | 1,682.8 | 2,643.2 |
| Total minutes on Zearn | 3,909,007 | 161.24 | 675.21 | 3.7 | 19.3 | 93.5 |

The table summarizes key educational metrics for Zearn teachers from July 2019 to June 2020.

### Study 1a: Identifying Key Teacher Behaviors Associated with Student Performance Using Independent Component Analysis

Zearn offers a variety of activities for students and teachers, resulting in a large number of usage variables. We chose to summarize this large set of online activities into more easily interpretable variables through a dimensionality reduction technique. We employed Independent Component Analysis (ICA) on these teacher behavioral variables, given the non-Gaussian nature of the data (Hyvärinen & Oja, 2000). This statistical approach allowed us to uncover latent variables that might have been obscured with traditional methods. The resulting components were weighted vectors of specific teacher activities and were estimated to maximize statistical independence in the data-generating process (Hyvärinen & Oja, 2000).

By examining the explained variance of each added component, we concluded that three independent components best portray the underlying dimensions of teacher behavior. (For more details, please refer to the SI section and Figure B.1). It is important to note that the results presented here reflect a revised version of our pre-experimental analysis. Although we maintained the same analytical framework, we addressed several inaccuracies identified in a post hoc review (for detailed information, please refer to the Supplementary Information). These corrections yielded more precise and robust results.

The significant finding from the ICA was the prominence of the first Independent Component (IC1), accounting for 15.39% of the variance in teacher Zearn activity, as indicated in Figure 3.1. This result is significant in nudge engineering, highlighting the need to focus interventions on elements encapsulated by IC1. Conversely, IC2 and IC3, with 12.56% and 6.3% variance explained, play lesser but still noteworthy roles.

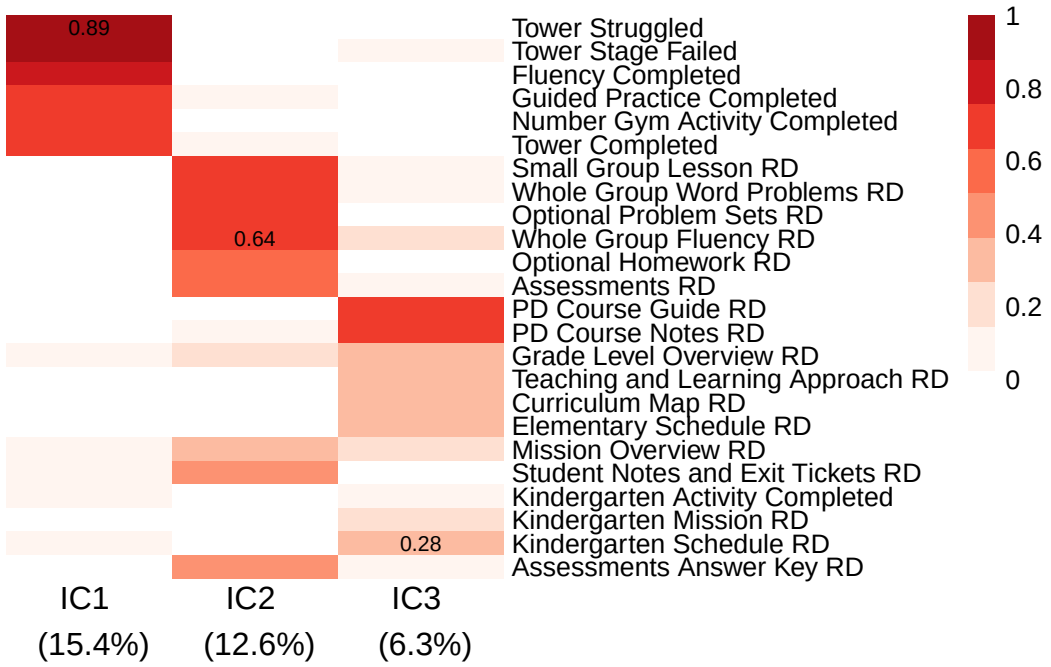



Figure 3.1: Independent Component Analysis (ICA) results.

The heatmap displays teacher actions in each row, while the columns represent the three independent components (IC1, IC2, IC3) that explained the most variance in the teacher behavioral data, with their respective percentage of variance explained in parentheses. The color gradient on the heatmap indicates the relative importance of each activity within these components. Note that these metrics pertain to teacher activity on the platform, not student actions. RD = Resource Download.

The activities with the highest weighting in IC1 included metrics associated with the Tower of Power (Struggled, Failed, and Completed) and other problem-solving exercises (i.e., Fluency exercises, Guided Practice, and Number Gym Activity) . Notably, the top two variables in this component involved an interactive feature of the Tower of Power when students struggle to understand a concept (i.e., Tower Struggled or Failed). The platform offers tailored support through Boosts, which break down questions into smaller steps, helping students understand and correct their mistakes. Given that the metrics in the ICs pertain to teacher activity, IC1 suggests that teachers proactively engage with these problem-solving activities, sometimes deliberately making mistakes, to better understand the student experience. This interaction may help teachers devise strategies to break down complex problems into simpler steps, known as instructional scaffolding (Beed et al., 1991; Cai et al., 2022).

After consulting with Zearn educators and administrators (see SI for details), it was that teachers with high levels of IC1 apply an empathy-driven pedagogy. In

this context, *empathy* refers to teachers' ability to comprehend and engage with students' challenges, as demonstrated by their focused attention on areas where students struggled or succeeded. The weightings in IC1 for activities like Tower Struggles (0.895), Fluency Completed (0.838), Guided Practice Completed (0.706), and Number Gym Activity Completed (0.661) were significantly higher compared to other activities. This pattern underlines this empathy-driven process in teaching, where teachers' engagement with the platform is focused on understanding student challenges.

The second component, IC2, showed substantial weightings on a variety of Resource Downloads (RD), particularly Small Group Lessons (0.717) and Whole Group (i.e., entire class) Word Problems (0.669). We labeled this component as "Classroom Activities." IC3, with strong weights on Resource Download (RD) for the Professional Development (PD) course guide (0.663) and course notes (0.657), was labeled "Professional Development." This component, accounting for a smaller variance (6.296%), was harder to interpret due to more diverse activities.

Table 3.2: Regression of log badges on the independent components.

| | **ln(Badges + 1)** | |
|---|---|---|
| | **All Schools** | **Zearn Curriculum** |
| IC 1 | 0.010* | 0.070*** |
| | (0.0043) | (0.0070) |
| IC 2 | 0.056*** | 0.056*** |
| | (0.0037) | (0.0040) |
| IC 3 | -0.006*** | -0.006** |
| | (0.0017) | (0.0023) |
| No. of Classes | -0.067*** | -0.038*** |
| | (0.0079) | (0.0100) |
| Average Intercept | 1.276 | 1.133 |
| R-Squared | 0.012 | 0.027 |
| Teachers | 2506 | 1413 |
| Classes | 4094 | 2389 |
| Weeks | 39 | 39 |
| Total | 115532 | 67549 |

**Note:** * p<0.05; ** p<0.01; *** p<0.001

The table displays the results of a panel regression model with clustered standard errors at the teacher level in parentheses.

Building upon these findings, we advanced to a fixed-effects panel regression model that accounted for various temporal and subject-specific variables and the potential impact of teachers handling multiple classes. The regression formula incorporated changes in independent components (IC1, IC2, IC3) as predictors for the change in the logarithm of badges (+1), accounting for individual teachers, classes, and weeks. As presented in Table 3.2, the regression highlighted a strong positive contemporaneous correlation between IC1 and badges, with a coefficient of 0.0101 ($p = 0.0186$), suggesting a significant impact on student achievement. Upon the recommendation of Zearn administrators, a supplementary analysis was conducted, focused exclusively on schools using Zearn either as their core curriculum or a key supplementary resource, effectively excluding teachers who independently chose Zearn despite the absence of school-wide implementation. This subset analysis revealed an effect of 0.07 ($p < .001$), indicating that IC1 is especially significant for those schools.

This correlation also has practical significance. It indicates that an increase in activities associated with IC1, such as increasing the encounters of "Tower Struggle" by one standard deviation, is associated with an approximate 1.1 percent increase in student badges. Although this increase may seem small, it may be substantial in the context of nudges, where even modest changes often lead to far-reaching effects.

### Study 1b: Identifying Teacher Calendar Predictability Associated with Student Performance

Prompted by Buyalskaya et al. (2023), we sought to uncover the subtleties of predictable behaviors within an educational setting. Our primary goal was to understand how regular teacher interactions with the Zearn platform impacted student learning outcomes. To measure this, we used the log-transformed average weekly badges per student over the entire school year as our dependent variable. We constructed our explanatory variables with careful consideration of the patterns that could identify predictable engagement and their relationship to student performance:

1. Login Percentages across months and days of the week: The frequency of logins across different time periods. For example, among all the logins for a teacher, we assess how many are from January or Tuesdays, compared to all other months and days of the week, respectively.

2. Average Minutes: Teachers' weekly average time spent on the platform, measured in minutes.

3. Average Streak: Average number of consecutive days in which the teacher logs in, excluding weekends.

4. Average Weekday Streak: Average number of consecutive weekdays in which the teacher logs in (e.g., three Tuesday in a row).

5. Average Days Between Logins: Average number of days between two login instances.

6. Total Logins: Total number of teacher logins over the entire school year.

To account for any school-specific factors that may have influenced the relationship between teacher behavior and student achievement, we estimated a linear regression model. Unlike Study 1, this regression did not follow each teacher or class across weeks, as our unit of analysis was a teacher summed across classrooms and averaged across all weeks. Note that our regression omitted the login percentages from July and Sunday, periods with low login incidence, to avoid multicollinearity.

The regression results, as detailed in Table 3.3, revealed that all weekday login percentages positively affected student badges. However, Friday logins stood out significantly, suggesting that specific days of the week have more influence on habitual engagement.

In particular, our analysis indicates that shifting 10% of logins from other weekdays to Fridays, without increasing overall platform usage, could boost student lesson completion by around 9.68%. For the typical teacher, this means switching from one Friday login per month to two while reducing one login from another weekday during that month, resulting in an increase of 13.78% in average weekly lesson completion. This outcome suggests an importance of strategic engagement rather than just login frequency.

Table 3.3: Regression of log average badges on calendar predictability variables.

| | Avg. Weekly Badges | |
|---|---|---|
| | **Estimate (SE)** | **P-value** |
| (Intercept) | 4.513 (2.759) | .102 |
| %Monday | 0.859 (0.567) | .13 |
| %Tuesday | 0.725 (0.449) | .107 |
| %Wednesday | 0.866 (0.628) | .168 |
| %Thursday | 0.831 (0.515) | .107 |
| %Friday | 1.788 (0.507) | <.001 |
| %Saturday | 0.938 (1.167) | .421 |
| %January | -4.118 (2.742) | .133 |
| %February | -2.803 (2.843) | .324 |
| %March | -3.818 (2.764) | .167 |
| %April | -3.858 (2.789) | .167 |
| %May | -3.193 (2.790) | .252 |
| %June | -2.662 (2.965) | .369 |
| %August | -4.008 (2.796) | .152 |
| %September | -4.117 (2.783) | .139 |
| %October | -3.757 (2.864) | .19 |
| %November | -3.989 (2.785) | .152 |
| %December | -3.185 (2.904) | .273 |
| Avg. Minutes | 0.011 (0.002) | <.001 |
| Avg. Streak | 0.061 (0.108) | .573 |
| Avg. Days Between Logins | 0.002 (0.003) | .435 |
| Avg. Weekday Streak | -0.132 (0.185) | .475 |
| Total Logins | 0.011 (0.002) | <.001 |
| R-squared | 0.133 | |
| N | 4273 | |

The table displays the results of the regression model with standard errors clustered at the school level in parentheses. Coefficients on months and days measure the difference between the effects of login percentage relative to the July percentage (for months) and the Sunday percentage (for days).

Based on our data analysis and focus group discussions with teachers and administrators, we hypothesized that this pronounced effect may be due to two key factors. First, Friday logins could facilitate "Reflective Catch-Up," enabling teachers to review and analyze the previous week's activities and make necessary adjustments. Second, "Foresighted Planning" may occur on Fridays as teachers proactively plan for the upcoming week, a practice less common on weekends.

**Study 2: Nudge Engineering—From Data to Intervention Design**

In Study 1, we discovered that specific psychopedagogical strategies, habits, and timing of teacher interactions were associated with promoting student success on digital platforms. With this foundation, we could craft effective educational interventions and strategies. By leveraging the insights from the first study, we aimed to develop targeted interventions to increase student learning outcomes.

**Study 2a: Using ICA Insights to Design an Empathy-Driven Teacher Intervention**

We used the first component from the ICA in Study 1a to design an empathy nudge. Our analysis in Study 1a indicated that this process was linked to the highest weighted behaviors in IC1 (note that the original analysis yielded an empathy coefficient of 0.0255 for the whole sample and 0.1434 for the curriculum-only subsample, versus 0.0101 and 0.07, respectively, in the revised analysis; refer to SI for details). This intervention involved sending emails to teachers, encouraging them to adopt a more student-centered teaching approach by viewing math problems from their students' perspective. The emails contained key related messages that mentioned the general findings of a "recent analysis" (i.e., Study 1a), emphasized the importance of empathy in teaching math, and included advice from other teachers and helpful tips for assisting students who struggle with a lesson. The emails also suggested specific actions, including using Zearn's features to view lessons from a student's perspective and checking the Tower Alerts Report (see SI for the complete emails). We hypothesized that this empathy approach would significantly enhance student performance.

**Study 2b: Using Calendar Predictability Insights to Design a Friday Login Intervention**

In Study 2b, we aimed to test whether nudging teachers to log in on Fridays, as opposed to an unspecified day, would improve student performance. Our approach involved sending emails to teachers, reminding them to log in on Fridays and suggesting reflective and planning behaviors to engage in during those sessions. These emails included testimonials from other teachers, research insights, and motivational messages to encourage habit formation. Teachers were also encouraged to review student progress and identify areas where students needed additional support

(see SI for the complete emails). Our rationale was that Friday logins would aid in reflecting on the week's activities and proactive planning for the following week.

In addition, we designed a control email tailored to this treatment. These emails were sent every Wednesday to remind Zearn teachers to review their Zearn Pace Report without the personalized behavioral prompts or motivational materials in the Friday treatment group emails. Although the control emails prompted participation with Zearn, they did not provide information on Friday logins, introspection, or strategic preparation.

We hypothesized that the Friday approach to teacher engagement with the platform would yield a greater impact on student achievement than the Wednesday control or our study-wide control.

### Intervention Impact Evaluation

In collaboration with Zearn, Study 2 became part of a large multi-arm *megastudy*, set in motion during a critical four-week period in 2021, involving over 140,000 teachers and nearly 3 million students. Teachers in our study taught a median of 20 students (mean = 21.30, SD = 15.31) in a median of 1 classroom (mean = 1.15, SD = 0.59). Before the intervention, teachers in our study had, on average, logged into Zearn a total of 3.61 times between July 1, 2021, and September 14, 2021 (SD = 8.80) (see SI for a complete description of the sample).

Table 3.4: Efficacy of different teacher engagement interventions on student learning outcomes.

| | Total Badges | |
|---|---|---|
| **Treatment** | **Estimate (SE)** | **P-value** |
| Empathy | 0.063 (0.029) | .0333 |
| Friday | 0.067 (0.029) | .0213 |
| Friday Control | 0.090 (0.029) | .00194 |
| Intercept (Study Control) | 0.097 (0.015) | <.001 |
| R-Squared | 0.685 | |
| N | 140461 | |

The table showcases the marginal effects of the "Empathy" and "Friday Login" interventions compared to the study-wide control as a baseline. It also includes the results from the Friday-specific control group. We measure student achievement by the number of badges students earned.

As stated in our pre-registration, we evaluated the impact of intervention messages on the number of math lessons completed by students during the four-week intervention period (Gallo et al., 2022a, 2022b). Students in the megastudy control condition completed a regression-estimated 1.780 lessons during the 4-week intervention period. Table 3.4 shows that our interventions increased the number of math lessons completed by students during the intervention period by a regression-estimated average of 0.0547 lessons, which is a 3.07% increase over the megastudy control condition. Specifically, the empathy treatment increased the number of lessons completed by students by 0.0628 lessons, or a 3.53% increase over the megastudy control condition (d = 0.0147, p = .018). The Friday treatment increased the number of lessons completed by students by 0.0550 lessons, or a 3.09% increase over the megastudy control condition (d = 0.0250, p = .068). The Friday treatment was not significantly different from the Friday-specific control (F(1,118137) = 0.39, p = .733).

Table S3 presents unstandardized coefficients from our primary regression analysis and unadjusted, robust SEs and CIs. Additionally, we used the Benjamini-Hochberg (BH) procedure to compute adjusted p-values, which help to control for false discovery rates when conducting multiple comparisons (Benjamini & Hochberg, 1995). Before adjusting for multiple hypothesis testing, all treatments exhibited significant benefits. However, only our treatment-specific control had a BH-adjusted p-value of less than 0.05. This intervention involved encouraging teachers to log in to Zearn weekly to receive updated student performance reports. Although reliable, the effect of this intervention was small, resulting in an estimated 0.0900 extra lessons completed in four weeks, or a 5.06% increase over the megastudy control (d = 0.0252, p = .002). Even after applying the James-Stein shrinkage procedure to adjust for the winner's curse (i.e., the maximum of 15 estimated effects is upward biased), we estimated that this intervention still produced 0.059 extra lessons completed, or a 3.30% increase over the control condition (James & Stein, 1992).

## 3.4 Discussion

This study aimed to improve student math learning on the Zearn platform by integrating data analysis into educational intervention. We identified critical teacher behaviors influencing student performance and evaluated two novel interventions: empathy and Friday logins.

Our first study revealed subtle but significant patterns in teacher engagement that traditional analyses might overlook. Study 1a identified a significant independent component strongly associated with metrics related to struggles and achievements, suggesting teachers' empathy-driven engagement, focusing on areas where students faced challenges or succeeded. This pattern also correlated with a significant increase in student achievement, aligning with findings emphasizing the importance of teacher empathy in educational outcomes (Battey et al., 2016; Hill et al., 2008). In Study 1b, we discovered that teachers who logged into Zearn on Fridays had a notable impact on student math performance. This behavior suggested a commitment to continuous planning and support, echoing the findings of Blanchard et al. (2016) that technology integration does not need large-scale changes in practices to enhance student learning. Overall, our research highlights the importance of teacher engagement and personalized instruction and feedback in improving student performance on the Zearn platform. As Ertmer et al. (2012) have emphasized, aligning student-centered beliefs and practices is vital to success, regardless of technological, administrative, or assessment barriers.

In Study 2a, the success of the empathy intervention can be attributed to its alignment with psychological principles that emphasize the importance of emotional connectivity between teachers and students. This intervention appears to have fostered a more engaging and supportive learning environment, a feature essential for the success of digital platforms (Koch et al., 2015; Lavecchia et al., 2016). In contrast, our initial hypothesis suggested that Friday logins might have a special effect on student outcomes. Study 2b's results highlighted that there were no statistically significant differences between the Friday treatment and its control (Wednesday). It is likely that the initial correlation between Friday and student achievement was spurious, or that no causal relationship exists between Friday logins and student achievement.

Notably, our study-specific control outperformed all other megastudy interventions. Initial analyses from Duckworth et al. (2024) suggest that this effect is due to the higher salience of personalization present, such as the suggestion of classroom-specific actions (e.g., "CLICK HERE to see which of your students are struggling"). The lack of additional content in this control highlighted actionable steps to engage with students, an effect which, in retrospect, aligns with previous literature.

Our study's insights transcend the immediate context. It showcases the potential of data-driven interventions to create strategies that cater to the unique needs of teachers and students. This approach paves the way for personalized and responsive

pedagogy. In the realm of digital learning, our findings underscore the pivotal role of teacher engagement and tailored content, which are crucial for replicating and enhancing the benefits of traditional classrooms. In essence, our research provides a roadmap for designing online educational tools that are more effective and engaging.

This innovative approach, while promising, is not without limitations. Our study identified behavioral patterns that resulted in both effective (i.e., the empathy treatment) and ineffective (i.e., the Friday treatment) interventions. This inconsistency highlights the inherent challenges in applying big data analysis and machine learning to complex contexts. These methods can occasionally identify spurious correlations that fail to translate into effective interventions. Further, the methods may not capture all of the complexities of behavior change: in our case, it is conceivable that an undetected confounding factor influenced both Friday logins and student achievement, which our treatment did not differentially modulate between Wednesday and Friday. Future studies are needed to develop more sophisticated approaches for pre-selecting and validating patterns from big data analyses to filter out potentially spurious correlations.

Our study is also limited by its focus on a specific demographic and educational context within Zearn. The data provide teacher and classroom interactions only within the online platform, which may reflect or influence in-class dynamics but does not directly measure in-class interactions. Consequently, generalizing our results to other educational settings, cultures, or age groups may be challenging. Future research could explore the relationship between online engagement patterns and in-class teaching practices. Further, we could not examine variation in performance among individual students within each classroom because of data aggregation, and future research could examine the variation in treatment effects among students within each classroom.

Additionally, while our approach was more cost-effective and less time-intensive, it achieved a more modest impact than the substantial effects seen in more intensive programs (Banerjee et al., 2007; Di Pietro, 2023). The simplicity of our email interventions and the short duration of the study likely contributed to these results, although their magnitude aligns with other reports from educational technology applications (Cheung & Slavin, 2013). Future research could explore more engaging and intensive intervention methods over extended periods to potentially yield greater impacts on learning outcomes.

While rooted in education data, our research introduces a paradigm with significant implications beyond its primary focus. By combining data-driven analysis with targeted behavioral interventions, our approach offers a flexible framework that can be adapted to various fields. Whether in healthcare, environmental behavior, or organizational management, our methodology demonstrates the potential to harness data insights for effective behavioral change. Hence, our study also serves as a catalyst for innovative approaches in diverse fields where behavior modification is crucial.

## 3.5 Materials and Methods

### Study 1

### Data Collection

The dataset used in this study was automatically collected by Zearn's servers during the 2019-2020 academic year (September 2019 to May 2020). The platform tracked user interactions, such as teacher and student logins, completed lessons by students, and professional development modules completed by teachers. Teacher actions were timestamped to the second, providing granular data on their behavior. To protect student privacy, student data was aggregated at the classroom-week level, including measures of student achievement and indicators of student struggles. We merged these data with a version of teacher data aggregated to the weekly level.

This study was conducted in accordance with ethical standards and received exempt status from the Institutional Review Board (IRB) at the University of Pennsylvania. The study's methodologies were designed to ensure the confidentiality and anonymity of all participants involved, adhering strictly to ethical guidelines for educational research.

To promote transparency and replicability of our study, we deposited the de-identified dataset and code used in our analyses in a publicly accessible database, available at the GitHub repository: https://github.com/SeanHu0727/zearn_nudge.git

### Inclusion Criteria

The dataset included various schools across Louisiana (see Fig S1 for geographic distribution). We also excluded inactive teachers (those with no recorded activity for over two months) from the dataset. We defined the following inclusion criteria for classrooms:

1. Classrooms linked to a single teacher
2. Classrooms with no more than seven months of inactivity during the academic year
3. Classrooms with an average of no less than five actively engaged students

In study 1a, we further categorized Zearn usage into curriculum and non-curriculum cases. *Curriculum* refers to scenarios where Zearn is integrated as a core component of the school's daily schedule. *Non-curriculum* cases, meanwhile, involve Zearn being used alongside different core curricula, resulting in varied consistency in usage.

### Analysis of Study 1a

### Independent Component Analysis (ICA)

We extracted all teacher behavioral variables from the dataset that displayed non-zero variance and standardized them to have a zero mean and unit variance. To determine the ideal number of independent components, we performed ICA using a range of components from 1 to 10. Our decision on the optimal number was informed by recognizing the "elbow" on the scree plot (i.e., the point at which adding more components yields diminishing increases in total explained variance), yielding three independent components. We used the `icafast` function from the R `ica` package for all ICAs conducted (Helwig, 2022).

### Panel Regression

We estimated a fixed-effects panel regression model with the `plm` package (Croissant & Millo, 2008) in R. The dependent variable was defined as:

$$\begin{aligned}\ln(\text{Badges}_{itc} + 1) = \beta_1 \text{IC1}_{it} + \beta_2 \text{IC2}_{it} + \beta_3 \text{IC3}_{it} \\ + \text{FE}_{\text{Teacher},i} + \text{FE}_{\text{Week},t} + \text{FE}_{\text{Class},c} + \epsilon_{itc}\end{aligned}$$

where $i$, $c$, and $t$ index the teacher, class, and week, respectively. The model includes fixed effects for teacher ($\text{FE}_{\text{Teacher},i}$), week ($\text{FE}_{\text{Week},t}$), and class ($\text{FE}_{\text{Class},c}$). Standard errors were clustered at the teacher level, calculated using the `vcovHC` function (Millo, 2017) in R.

### Study 1b

### Linear Regression

We estimated a linear model using the `lm` function (Team, 2023) in R:

$$\begin{aligned}\sum_{c=1}^{C_i} (\text{Avg. Badges})_c = \beta_0 + \sum_{m=1,m\neq 7}^{12} \beta_m \text{Login\%}_{m,i} + \sum_{d=1,d\neq 7}^{7} \beta_{12+d} \text{Login\%}_{d,i} \\ + \beta_{19}(\text{Avg. Minutes})_i + \beta_{20}(\text{Avg. Streak})_i \\ + \beta_{21}(\text{Avg. Days Between Logins})_i + \text{FE}_{\text{School}} + \epsilon_i\end{aligned}$$

for teacher $i$ with $C_i$ classes. `Login%` represents the percentage of logins by teacher $i$ during each month $m$ and on each day of the week $d$ relative to other months and days, respectively. We exclude July and Sunday to prevent multicollinearity. Standard errors are clustered by school with the `vcovCL` function (Zeileis, 2004; Zeileis et al., 2020) in R.

### Study 2

We used the findings from Study 1 to inform the creation of two interventions as part of a larger multi-arm megastudy that involved 15 sets of nudges.

### Implementation

Study 2 was conducted in collaboration with Zearn ("Zearn Math: Top-Rated Math Learning Platform", 2023) and was pre-registered for the fall of 2021. To incentivize teacher participation during our intervention period from September 15, 2021, to October 12, 2021, all teachers on the platform received two messages on September 1 and 8, 2021. These messages informed them they had been enrolled in the "Zearn Math Giveaway" and that every email opened until October 12, 2021, would earn them tickets. These tickets were used to enter drawings for various prizes, such as autographed children's books, stickers, and gift cards.

### Data Preprocessing

Following the megastudy's pre-registered analysis plan (https://osf.io/dgpkn), we restricted analyses to Zearn Math teachers who were assigned to one of the megastudy's conditions and who taught in at least one classroom with at least one student. However, our analyses excluded teachers who: (1) did not receive any emails because

they had inactive accounts, invalid email addresses, or had opted out of receiving messages (n = 133,722 teachers), (2) neither logged onto the Zearn Math platform nor had an associated student who logged on the platform between March 1, 2021 and September 14, 2021 (n = 126,856 teachers), (3) had more than 150 students associated with their Zearn Math account as of October 18, 2021 (n = 6,766 teachers), or (4) had more than 6 classes associated with their Zearn Math account as of October 18, 2021 (n = 6,675 teachers). Among Zearn Math classrooms associated with the remaining 149,097 teachers, we further excluded: (5) 9,143 classrooms associated with more than one Zearn Math teacher (n = 12,012 teachers) and (6) 346 classrooms with grade levels corresponding to high school or post-high school (n = 141 teachers).

After exclusions, we randomized $N = 140,461$ teachers across 22,281 schools who served 2,992,077 students in 161,722 classrooms into one of the intervention conditions or the control condition ($N_{\text{control}} = 29,513$). The control condition was larger than the interventions to account for multiple comparisons. Among $n = 16,372$, or 11.66%, of teachers, at least one of two problems occurred in the emails sent by Zearn Math during the intervention period: an email message that was intended but not sent ($n = 13,568$, or 9.66% of teachers), or an email message that was sent but not intended (i.e., from a different treatment condition; $n = 2,804$, or 2.00% of teachers). As these email problems were systematically related to treatment assignment ($\chi^2 = 33.01$, $df = 15$, $p = .005$), we did not exclude these participants and conducted intent-to-treat analyses. Refer to the SI for email problem prevalence by condition and study analyses that exclude or adjust for email problems, respectively.

**Statistical Analysis**

We followed our pre-registered analysis plan to assess the effect of each treatment on the primary outcome of interest: math lessons completed by students during our four-week intervention period (Gallo et al., 2022a, 2022b). We estimated a weighted ordinary least squares (OLS) regression with the `areg` command in Stata (StataCorp, 2023). Each teacher's observations were weighted proportionally to the total number of students in their Zearn classroom(s).

The primary predictors were indicators for each intervention, omitting the control condition. The regression also included the following control variables: (1) school fixed effects, (2) an indicator for the teacher's account type (free or paid), (3) the

number of times the teacher logged into Zearn prior to the study, from August 1 to September 14, 2021, (4) the total number of students in the teacher's classroom(s) as of October 18, 2021, (5) the number of classrooms associated with the teacher as of October 18, 2021, (6) the number of days since the teacher obtained a Zearn account prior to the study's launch, (7) the number of days separating the study's launch and the start of the teacher's school year, (8) the average number of lessons completed by a teacher's students from the start of their school year to the start of the intervention (or from July 14, 2021, if the school year start was not known), (9) whether the teacher opened our September 1, 2021 email announcing the upcoming Zearn Math Giveaway, (10) a similar indicator for our September 8, 2021, email reminding them of the giveaway, and (11) the percentage of a teacher's students in each grade except for third grade to avoid multicollinearity, since for most teachers, students were in a single grade.

### Study 2a: Empathy Intervention Design and Implementation

We designed four emails encouraging teachers to adopt a more student-centered perspective when engaging with the Zearn platform to gain pedagogical content knowledge. The emails included testimonials from experienced teachers, research-based insights on the role of empathy in math education, and specific strategies for using Zearn's features to understand and address student challenges. A population of 7,443 teachers was chosen at random to receive these emails.

### Study 2b: Friday Logins Intervention Design and Implementation

We designed four emails encouraging teachers to regularly log into the Zearn platform on Fridays. The emails emphasized the benefits of using Fridays for reflective review and proactive planning, highlighting how this practice could help teachers better support student learning. Teachers were provided with specific suggestions for activities during these Friday sessions, such as analyzing student progress data, identifying areas for improvement, and planning targeted interventions for the upcoming week. A total of 7,476 teachers was randomly selected to receive these emails. Additionally, to assess the unique impact of the Friday login habit, we included an active control condition that received similar email prompts on Wednesdays with links to specific actions on Zearn but without a focus on Fridays. A group of 7,577 teachers were randomly chosen to participate in this control group.

## 3.6 Acknowledgments

We thank Billy McRae, Michael Irvine, and Audrieanna Burgin for their valuable insights and support during our data analysis. We thank Zearn administrators and teachers for feedback on our intervention design. We also thank Cassandra Horri for the helpful comments that improved this manuscript.

## References

Alabdulaziz, M. S. (2021). Covid-19 and the use of digital technology in mathematics education [31 citations (Crossref) [2024-01-03]]. *Education and Information Technologies*, *26*(6), 7609–7633. https://doi.org/10.1007/s10639-021-10602-3

Al-Shabandar, R., Hussain, A. J., Liatsis, P., & Keight, R. (2018). Analyzing learners behavior in moocs: An examination of performance and motivation using a data-driven approach [29 citations (Crossref) [2024-01-03]]. *IEEE Access*, *6*, 73669–73685. https://doi.org/10.1109/access.2018.2876755

Banerjee, A. V., Cole, S., Duflo, E., & Linden, L. (2007). Remedying education: Evidence from two randomized experiments in india [599 citations (Crossref) [2024-01-03]]. *The Quarterly Journal of Economics*, *122*(3), 1235–1264. https://doi.org/10.1162/qjec.122.3.1235

Battey, D., Neal, R. A., Leyva, L., & Adams-Wiggins, K. (2016). The interconnectedness of relational and content dimensions of quality instruction: Supportive teacher–student relationships in urban elementary mathematics classrooms [18 citations (Crossref) [2024-01-03]]. *The Journal of Mathematical Behavior*, *42*, 1–19. https://doi.org/10.1016/j.jmathb.2016.01.001

Beed, P. L., Hawkins, E. M., & Roller, C. M. (1991). Moving learners toward independence: The power of scaffolded instruction [Publisher: [Wiley, International Reading Association]]. *The Reading Teacher*, *44*(9), 648–655. http://www.jstor.org/stable/20200767

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing [22907 citations (Crossref) [2024-01-03]]. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Blanchard, M. R., LePrevost, C. E., Tolin, A. D., & Gutierrez, K. S. (2016). Investigating technology-enhanced teacher professional development in rural, high-poverty middle schools [57 citations (Crossref) [2024-01-03]]. *Educational Researcher*, *45*(3), 207–220. https://doi.org/10.3102/0013189x16644602

Buyalskaya, A., Ho, H., Milkman, K. L., Li, X., Duckworth, A. L., & Camerer, C. (2023). What can machine learning teach us about habit formation? Evidence from exercise and hygiene. *Proceedings of the National Academy of Sciences*, *120*(17), e2216115120. https://doi.org/10.1073/pnas.2216115120

Cai, Z., Mao, P., Wang, D., He, J., Chen, X., & Fan, X. (2022). Effects of scaffolding in digital game-based learning on student's achievement: A three-level meta-analysis [35 citations (Crossref) [2024-06-04]]. *Educational Psychology Review*, *34*(2), 537–574. https://doi.org/10.1007/s10648-021-09655-0

Cheung, A. C., & Slavin, R. E. (2013). The effectiveness of educational technology applications for enhancing mathematics achievement in k-12 classrooms: A meta-analysis. *Educational Research Review*, *9*, 88–113. https://doi.org/10.1016/j.edurev.2013.01.001

Chiang, P.-J., Lin, Y.-W., & Tseng, C.-L. (2016). The effect of blended learning in mathematics course [27 citations (Crossref) [2024-01-03]]. *EURASIA Journal of Mathematics, Science and Technology Education*, *13*(3). https://doi.org/10.12973/eurasia.2017.00641a

Croissant, Y., & Millo, G. (2008). Panel data econometrics in *r* : The **plm** package. *Journal of Statistical Software*, *27*(2). https://doi.org/10.18637/jss.v027.i02

Di Pietro, G. (2023). The impact of COVID-19 on student achievement: Evidence from a recent meta-analysis. *Educational Research Review*, *39*, 100530. https://doi.org/10.1016/j.edurev.2023.100530

Ertmer, P. A., Ottenbreit-Leftwich, A. T., Sadik, O., Sendurur, E., & Sendurur, P. (2012). Teacher beliefs and technology integration practices: A critical relationship [844 citations (Crossref) [2024-01-03]]. *Computers & Education*, *59*(2), 423–435. https://doi.org/10.1016/j.compedu.2012.02.001

Ewing, E. L., & Green, T. L. (2021). Beyond the headlines: Trends and future directions in the school closure literature [5 citations (Crossref) [2024-01-03]]. *Educational Researcher*, *51*(1), 58–65. https://doi.org/10.3102/0013189x211050944

Fahle, E. M., Kane, T. J., Patterson, T., Reardon, S. F., Staiger, D. O., & Stuart, E. A. (2023, May). *School district and community factors associated with learning loss during the covid-19 pandemic* (tech. rep.). Cambridge, MA.

Gallo, M., Camerer, C., Manning, B., Milkman, K., & Duckworth, A. (2022a). Zearn: Empathy study [Publisher: Open Science Framework]. https://doi.org/10.17605/OSF.IO/KDP2U

Gallo, M., Camerer, C., Manning, B., Milkman, K., & Duckworth, A. (2022b). Zearn: Nudging weekly logins [Publisher: Open Science Framework]. https://doi.org/10.17605/OSF.IO/JFV3B

Greene, J. A. (2022). What can educational psychology learn from, and contribute to, theory development scholarship? [9 citations (Crossref) [2024-01-02]]. *Educational Psychology Review*, *34*(4), 3011–3035. https://doi.org/10.1007/s10648-022-09682-5

Haleva, L., Hershkovitz, A., & Tabach, M. (2020). Students' activity in an online learning environment for mathematics: The role of thinking levels [7 citations (Crossref) [2024-01-03]]. *Journal of Educational Computing Research*, *59*(4), 686–712. https://doi.org/10.1177/0735633120972057

Hanushek, E. A. (2020). Education production functions [29 citations (Crossref) [2024-01-03]]. *The Economics of Education*, 161–170. https://doi.org/10.1016/b978-0-12-815391-8.00013-6

Helwig, N. E. (2022). *Ica: Independent component analysis* (tech. rep.). https://CRAN.R-project.org/package=ica

Hershcovits, H., Vilenchik, D., & Gal, K. (2020). Modeling engagement in self-directed learning systems using principal component analysis [8 citations (Crossref) [2024-01-03]]. *IEEE Transactions on Learning Technologies*, *13*(1), 164–171. https://doi.org/10.1109/tlt.2019.2922902

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study [497 citations (Crossref) [2024-01-03]]. *Cognition and Instruction*, *26*(4), 430–511. https://doi.org/10.1080/07370000802177235

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications [5771 citations (Crossref) [2023-12-27]]. *Neural Networks*, *13*(4-5), 411–430. https://doi.org/10.1016/S0893-6080(00)00026-5

James, W., & Stein, C. (1992). Estimation with quadratic loss [Series Title: Springer Series in Statistics DOI: 10.1007/978-1-4612-0919-5_30]. In S. Kotz & N. L. Johnson (Eds.). Springer New York. http://link.springer.com/10.1007/978-1-4612-0919-5_30

Koch, A., Nafziger, J., & Nielsen, H. S. (2015). Behavioral economics of education [103 citations (Crossref) [2024-01-03]]. *Journal of Economic Behavior & Organization*, *115*, 3–17. https://doi.org/10.1016/j.jebo.2014.09.005

Lavecchia, A., Liu, H., & Oreopoulos, P. (2016). Behavioral economics of education [47 citations (Crossref) [2024-01-03]]. *Handbook of the Economics of Education*, 1–74. https://doi.org/10.1016/b978-0-444-63459-7.00001-4

Li, Q., & Ma, X. (2010). A meta-analysis of the effects of computer technology on school students' mathematics learning [265 citations (Crossref) [2024-06-04]]. *Educational Psychology Review*, *22*(3), 215–243. https://doi.org/10.1007/s10648-010-9125-8

Li, S., & Wang, W. (2022). Effect of blended learning on student performance in k-12 settings: A meta-analysis [12 citations (Crossref) [2024-06-04]]. *Journal of Computer Assisted Learning*, *38*(5), 1254–1272. https://doi.org/10.1111/jcal.12696

Liljedahl, P., & Oesterle, S. (2020). Teacher beliefs, attitudes, and self-efficacy in mathematics education [2 citations (Crossref) [2024-01-03]]. *Encyclopedia of Mathematics Education*, 825–828. https://doi.org/10.1007/978-3-030-15789-0_149

Lynch, K., Hill, H. C., Gonzalez, K. E., & Pollard, C. (2019). Strengthening the research base that informs stem instructional improvement efforts: A meta-analysis [54 citations (Crossref) [2024-06-04]]. *Educational Evaluation and Policy Analysis*, *41*(3), 260–293. https://doi.org/10.3102/0162373719849044

Mawardi, D. N., Budiningsih, C. A., & Sugiman. (2023). Blended learning effect on mathematical skills: A meta-analysis study [0 citations (Crossref) [2024-06-04]]. *Ingénierie des systèmes d information*, *28*(1), 197–204. https://doi.org/10.18280/isi.280122

Meeter, M. (2021). Primary school mathematics during the covid-19 pandemic: No evidence of learning gaps in adaptive practicing results [31 citations (Crossref) [2024-01-03]]. *Trends in Neuroscience and Education*, *25*, 100163. https://doi.org/10.1016/j.tine.2021.100163

Millo, G. (2017). Robust standard error estimators for panel models: A unifying approach [45 citations (Crossref) [2023-12-30]]. *Journal of Statistical Software*, *82*(3). https://doi.org/10.18637/jss.v082.i03

Qiu, F., Zhang, G., Sheng, X., Jiang, L., Zhu, L., Xiang, Q., Jiang, B., & Chen, P.-k. (2022). Predicting students' performance in e-learning using learning process and behaviour data [35 citations (Crossref) [2024-01-03]]. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-021-03867-8

Ran, H., Kasli, M., & Secada, W. G. (2020). A meta-analysis on computer technology intervention effects on mathematics achievement for low-performing students in k-12 classrooms [21 citations (Crossref) [2024-01-03]]. *Journal of Educational Computing Research*, *59*(1), 119–153. https://doi.org/10.1177/0735633120952063

Ran, H., Kim, N. J., & Secada, W. G. (2021). A meta-analysis on the effects of technology's functions and roles on students' mathematics achievement in k-12 classrooms [11 citations (Crossref) [2024-01-03]]. *Journal of Computer Assisted Learning*, *38*(1), 258–284. https://doi.org/10.1111/jcal.12611

Sadaf, A., Wu, T., & Martin, F. (2021). Cognitive presence in online learning: A systematic review of empirical research from 2000 to 2019 [14 citations (Crossref) [2024-01-03]]. *Computers and Education Open*, *2*, 100050. https://doi.org/10.1016/j.caeo.2021.100050

Salazar, A., Serrano, A., & Vergara, L. (2007). Non-parametric ica reveals learning styles in education activities through the web [0 citations (Crossref) [2024-01-03]]. *2007 IEEE Workshop on Machine Learning for Signal Processing*. https://doi.org/10.1109/mlsp.2007.4414316

Shin, D., & Shim, J. (2020). A systematic review on data mining for mathematics and science education [22 citations (Crossref) [2024-01-03]]. *International Journal of Science and Mathematics Education*, *19*(4), 639–659. https://doi.org/10.1007/s10763-020-10085-7

StataCorp. (2023). *Stata statistical software*. StataCorp LLC.

Team, R. C. (2023). *R: A language and environment for statistical computing* (tech. rep.). Vienna, Austria. https://www.R-project.org/

Wang, X. S., Perry, L. B., Malpique, A., & Ide, T. (2023). Factors predicting mathematics achievement in pisa: A systematic review [2 citations (Crossref) [2024-01-02]]. *Large-scale Assessments in Education*, *11*(1), 24. https://doi.org/10.1186/s40536-023-00174-8

Young, J. R. (2016). Unpacking tpack in mathematics education research: A systematic review of meta-analyses [15 citations (Crossref) [2024-01-03]]. *International Journal of Educational Methodology*, *2*(1), 19–29. https://doi.org/10.12973/ijem.2.1.19

Yu, Q., Yu, K., Li, B., & Wang, Q. (2023). Effectiveness of blended learning on students' learning performance: A meta-analysis. *Journal of Research on Technology in Education*, 1–22. https://doi.org/10.1080/15391523.2023.2264984

Zearn math: Top-rated math learning platform. (2023). https://about.zearn.org/

Zeileis, A. (2004). Econometric computing with hc and hac covariance matrix estimators [581 citations (Crossref) [2024-01-31]]. *Journal of Statistical Software*, *11*(10). https://doi.org/10.18637/jss.v011.i10

Zeileis, A., Köll, S., & Graham, N. (2020). Various versatile variances: An object-oriented implementation of clustered covariances in *r* [232 citations (Crossref) [2024-01-31]]. *Journal of Statistical Software*, *95*(1). https://doi.org/10.18637/jss.v095.i01

Zierer, K. (2021). Effects of pandemic-related school closures on pupils' performance and learning in selected countries: A rapid review [57 citations (Crossref) [2024-01-03]]. *Education Sciences*, *11*(6), 252. https://doi.org/10.3390/educsci11060252

*Chapter 4*

# PERCEIVED WARMTH AND COMPETENCE PREDICT CALLBACK RATES IN META-ANALYZED NORTH AMERICAN LABOR MARKET EXPERIMENTS

Gallo, M., Hausladen, C. I., Hsu, M., Jenkins, A. C., Ona, V., & Camerer, C. F. (2024). Perceived warmth and competence predict callback rates in meta-analyzed North American labor market experiments (T. Otterbring, Ed.). *PLOS ONE*, *19*(7), e0304723. https://doi.org/10.1371/journal.pone.0304723

## 4.1 Abstract

We meta-analyzed 32 experimental *correspondence* studies investigating discrimination in the North American labor market. We collected data on 592 different social signals to examine the impact of social perception (warmth, competence) on callback rates. Our analysis found that social perception predicts callback rates for studies varying race and gender, as signaled by names. However, for studies varying other categories, such as sexuality and disability, the direction of social perception's impact on callback rates is less clear. These findings provide important insights into how social perception affects labor market outcomes and highlight areas where further research is needed.

## 4.2 Introduction

Discrimination is costly for organizations and detrimental to society. Labor market discrimination occurs when individuals are treated unequally based on their observable characteristics, even when those characteristics should not impact expected job performance. Despite increasing awareness of the advantages of diverse teams among employers and ongoing civil rights activism leading to legal protections against identity-based discrimination, people from marginalized groups still seem to face disparate treatment in the labor market. However, because employer subjective expectations of productivity are rarely observed, it is difficult to conclusively pinpoint specific instances of discrimination (Bertrand and Duflo, 2017; Bertrand and Mullainathan, 2004).

To try to control for subjective expectations, experimental *correspondence studies* (or *audit studies*) were developed starting in the 1960s (Daniel, 1968; Lippens et al.,

2023; Riach and Rich, 2002). Correspondence studies strive to control expectations by creating identical pairs of artificial resumes (with matching backgrounds, skills, education, etc.) and sending them to potential employers. Typically, only one categorical factor—such as race, gender, or sexuality—is varied between each matched resume pair. Everything else on the two resumes is the same. The test statistic is the difference in callbacks for the controlled variable. A callback refers to any response from the employer expressing interest in the candidate, in the form of a phone call or email (earlier studies also included letter replies)[1]. These studies have documented common patterns of discrimination across different social categories (although race is the strongest in the studies we include).

Our analysis explores the extent to which stereotyped responses to social groups, as identified in the correspondence labor market studies, are associated with social perceptions of those groups. Perceptions are measured in a two-dimensional space of warmth and competence based on extensive evidence that the two-variable warmth-competence reduction robustly explains a surprising amount of variation across perceptions and behavioral reactions to social categories. Warmth is the perception of how is good or bad another person's intentions are. Competence is how capable a person is of acting on their intentions (Fiske et al., 2007).

Emerging research is suggestive that stereotypes about warmth and competence may contribute to labor market discrimination (Agerström et al., 2012; Jenkins et al., 2018; Lippens et al., 2023; Veit and Thijsen, 2021); in particular, in a recent analysis (Jenkins et al., 2018), applicants whose racial group was associated with higher perceived warmth received significantly more callbacks based on data from two field studies. Although suggestive, this evidence comes from a few studies in a limited set of hiring domains.

For example, names can convey many attributes, including gender, ethnicity, and socioeconomic status. Crucially, when averaged across raters, warmth and competence scores for different groups are highly consistent across samples, suggesting that they reflect culturally shared stereotypes rather than idiosyncratic individual social perceptions (Jenkins et al., 2018).

Despite their success in documenting discrimination based on single social identities, these studies have limitations. For instance, they fail to account for intersectionality, where multiple social identities combine in complex and non-additive ways to influ-

[1]While phone callbacks may be perceived differently from written ones, most studies in our sample do not differentiate between these types.

ence treatment within the labor market (Browne and Misra, 2003; Nicolas and Fiske, 2023; Rosette et al., 2018; Thatcher et al., 2023). People with multiple marginalized identities are subjected to more frequent and severe workplace harassment (Hollis, 2018) and experience more obstacles to promotion (Bloch et al., 2020). Our analysis of previous studies is not able to measure intersectionality effects. To address this core limitation, we capture stereotype-based social perceptions that underlie the disparate treatment of different social groups.

We build on prior work with a comprehensive and comparative approach to 1) determine the relationship between warmth, competence, and callback rates across multiple studies, 2) evaluate the consistency and uniformity of the results, and 3) compare these effects across social categories, study designs, and industries. Our first contribution is demonstrating the effects of social perceptions on discrimination within studies that use names as a social signal for group membership. Second, we analyze the heterogeneity of this relationship with an extensive range of correspondence studies with non-name signals. Finally, we outline potential avenues for future investigation informed by our research findings, furthering our understanding of the interplay of psychological factors and discrimination within the labor market.

**Hypotheses**

We hypothesized that:

1. Average perceptions of warmth and competence would significantly predict callback rates in labor market experiments.

2. Both mean perceived warmth and competence would be positively correlated to callback rates.

## 4.3 Data

Figure 4.1 provides an overview of the correspondence studies we use, classifying them according to the investigated categories and the number of *signals* used to convey these traits. Typically, an applicant's group affiliation is not explicitly stated but subtly signaled through associations with 1) names, which are indicative of race, gender, or age, and 2) other characteristics (e.g., membership in a college LGBTQ club). Signals are chosen to maximally distinguish groups (e.g., Sarah Davis (white, female), Deshawn Jefferson (black, male)). Note that, unlike studies using names, category studies usually employ a limited number of distinct signals; therefore, we choose to analyze name and category studies separately.

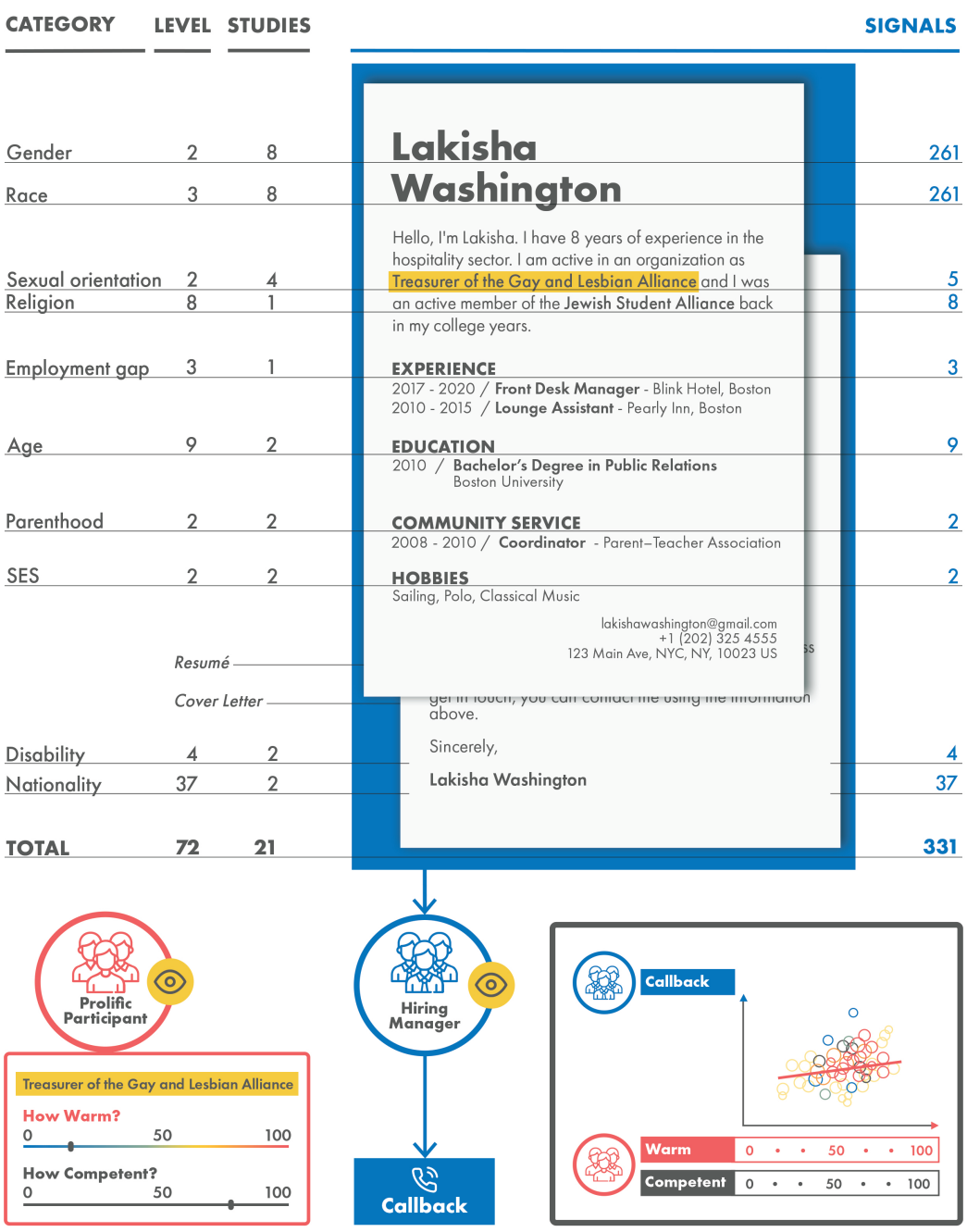



Figure 4.1: Graphical abstract

The total number of studies, categories, and signals included in our meta-analysis, along with our statistical estimation strategy. The numbers indicate the total counts of studies, categories, and signals. For a detailed overview of signals by category for which raw data was obtained, refer to Table S1. Please note that the total number of studies is 21, as some are included in more than one category. Example signals are presented in the middle column (the resume). The data sources are shown on the right-hand side: hiring managers made callback decisions based on resumes (in green). Separately, we collected warmth and competence ratings on prolific, where participants (in red) only saw the respective signal (indicated in yellow). Our estimation strategy is visualized in the gray box in the bottom right corner: we used the averages of warmth and competence ratings to predict the callback percentage.

## 4.4 Economic Models of Discrimination

Discrimination, particularly in the labor market, is a multifaceted phenomenon that can be understood through various theoretical lenses. Three primary theories have been proposed to explain the origins and mechanisms of discrimination: cultural, economic, and institutional models (Arrow, 1973; Becker, 1971; Brinton and Nee, 1998; McPherson et al., 2001; Phelps, 1972; Tajfel, 1974).

The cultural model of discrimination, often referred to as taste-based discrimination, was first proposed by economist Gary Becker in the 1950s. Becker argued that employers might have a distaste for ethnic minority groups, leading to economically inefficient hiring decisions based on their cultural preferences (Becker, 1971). This model suggests that employers are willing to pay a price for their preferences, such as higher wages for majority members. However, Becker's model only provides a framework to analyze the consequences of taste and does not explain where these preferences originate.

To understand the origins of taste, we turn to psychological approaches that explain the negative evaluation of others. These approaches include social identity theory (Tajfel, 1974) and the concept of homophily (McPherson et al., 2001). Becker's tastes may also be explained by what Bogardus, 1925, refers to as social distance, the perceived social distance varies across ethnic minority groups, resulting in an ethnic hierarchy (Bessudnov and Shcherbak, 2018; Hagendoorn, 1995; Hagendoorn and Hraba, 1989; Verkuyten et al., 1996).

The stereotype content model (SCM) argues that group stereotypes are a consequence of two interpersonal impressions: warmth and competence (Fiske et al., 2007). This model is particularly relevant to our study as it provides a comprehensive analysis of racial discrimination in the hiring process. The SCM is flexible to underlying theories of discrimination, offering a model-agnostic approach. This flexibility allows the SCM to capture various factors that drive discrimination, such as hostility toward foreign cultural norms and implicit discrimination.

The economic, or rational, model of discrimination, also known as statistical discrimination theory, postulates that employers act out of economic self-interest. Due to incomplete information and negative group beliefs about the skills of ethnic minorities, employers prefer majority candidates (Arrow, 1973; Phelps, 1972). This theory focuses on the individual decision-making of employers, suggesting that they use stereotypes about the productivity of social groups to make individual hiring decisions.

Similarly, the institutional model of discrimination emphasizes that the behavior of social actors is shaped by their social context (Brinton and Nee, 1998). This perspective suggests that employers' recruitment and reward behavior depends on the social norms, laws, and organizational practices that influence their decisions. In this view, the discriminatory behavior of employers is based on the systemic and often unconscious adherence to contextually derived biases (Di Stasio et al., 2019).

In conclusion, understanding discrimination in the labor market requires a multi-faceted approach that considers cultural, economic, and institutional factors. The cultural model, with its focus on taste-based discrimination and social distance, provides a framework for understanding the preferences that drive discriminatory behavior. The economic model, or statistical discrimination theory, offers insights into how employers' self-interest and stereotypes about productivity influence their hiring decisions. The institutional model emphasizes the role of social context in shaping employers' recruitment and reward behavior.

The Stereotype Content Model (SCM) is particularly relevant as it provides a comprehensive analysis of racial discrimination in the hiring process and is flexible to underlying theories of discrimination. This flexibility allows the SCM to capture various factors that drive discrimination, such as hostility toward foreign cultural norms and implicit discrimination.

Our study aims to capture these interactions and provide a comprehensive understanding of discrimination in the labor market. By focusing on the dimensions of warmth and competence, we hope to shed light on the stereotype-based social perceptions that underlie the disparate treatment of different social groups.

## 4.5 Stereotypes and Discrimination

This section lays the foundation for the present research by examining correspondence studies and their shortcomings by reviewing key studies that have examined stereotypes and discrimination in the context of employment.

Quillian and Lee (2023) examined racial discrimination after the callback stage and aimed to fill a gap in the literature by reviewing evidence from all available field experimental studies of racial or ethnic discrimination in hiring that proceed to the job offer outcome. The study's central idea is that there is considerable additional discrimination in hiring after the callback stage. Quillian et al. meta-analyzed 12 studies encompassing more than 13,000 job applications. The sample size was substantial, and the variables studied included the callback and job offer

outcomes. Analysis of the data revealed that majority applicants in the sample received 53% more callbacks than comparable minority applicants, but majority applicants received 145% more job offers than comparable minority applicants (Quillian & Lee, 2023).

Quillian and Lee (2023) concluded there is significant additional discrimination from interview to job offer, which is weakly correlated (r = 0.21) with the level of discrimination earlier in the hiring process. This finding highlights the extent of racial discrimination in the hiring process, even after the callback stage. These results suggest that statistical discrimination theory cannot adequately explain this discrimination since employers would have to be less likely to rely on group stereotypes as their information about an individual application increases. The study, however, did not explicitly address the reasons behind the additional discrimination after the callback stage, leaving a gap in understanding the underlying factors contributing to this discrimination.

Still, the Quillian and Lee (2023) study is relevant to the present study as it provides a comprehensive analysis of racial discrimination in the hiring process, an essential aspect of the present research. Quillian et al. also argued that economic models of discrimination fail to capture the various factors that drive discrimination (e.g., hostility toward foreign cultural norms) and implicit discrimination. Our use of social perception is immune to this effect, as it offers a model-agnostic approach to discrimination. Further, the article also revealed significant heterogeneity in how callback ratios differ across categories, which is an important finding in our study. Lastly, the Quillian and Lee (2023) study shows that callbacks (the dependent variable in our study) underestimate the total level of discrimination in the labor market, and they are only a lower bound of a much larger effect. This finding is consistent with our assumption that stereotypes are much more salient in person-to-person interactions such as interviews.

Lancee (2019) also reviewed experiments on ethnic discrimination in hiring processes across different contexts. The unifying question revolved around the extent of discrimination faced by ethnic minorities in the labor market and how it varies across different countries and ethnic groups. Lancee aimed to fill a gap in the literature by providing a cross-national perspective on this issue. The field experiment was conducted in five European countries: Germany, the Netherlands, Spain, Norway, and the United Kingdom. The methodology involved sending out fictitious job applications for nine occupations, with the applicant's ethnicity being the primary

variable (Lancee, 2019). The sample size was 19,181 job applications, and Lancee found that ethnic discrimination was prevalent in all countries studied, but the extent and the groups most affected vary.

Lancee (2019) concluded that ethnic hiring discrimination results from a complex interaction between institutions and social norms. For instance, Moroccan applicants were discriminated against in the Netherlands more often than in Spain, despite the latter having a high unemployment rate (Lancee, 2019). Lancee also found that discrimination is based not only on the applicant's country of origin but also on the specific labor market and the job characteristics in question. Lancee, however, did not fully explore the mechanisms behind the observed discrimination. While the study provides valuable insights into the existence and extent of ethnic discrimination in hiring, it does not delve deeply into why the same type of ethnic discrimination is more pronounced in certain countries. Our study aims to fill the gaps left by this research by examining the role of social perception in discrimination, which could be one of the mechanisms behind the observed ethnic discrimination in hiring. Our method is also flexible in relation to underlying theories of discrimination and could provide further insights into differential discrimination rates across contexts.

Researchers have also examined social perception in the hiring process. For example, Veit et al. (2021) investigated the role of social perception in hiring discrimination across five European countries. The research question focused on how warmth and competence influence hiring decisions. The study filled a gap in the literature on the impact of these perceptions on a global scale rather than focusing on a single country or region. The study was a harmonized cross-national field experiment involving 9,000 fictitious job applications sent to real job openings in nine countries. Veit et al. randomly assigned signals of warmth and competence by adding statements to the cover letter. Veit et al. (2021) found that manipulating competence significantly increases callbacks, while manipulating warmth has a less consistent impact.

Veit et al. (2021) concluded that hiring discrimination based on social perception is a global phenomenon. The study emphasizes the importance of perceived competence in hiring decisions, suggesting that employers may prioritize competence over warmth when making hiring decisions. Veit et al.'s study contributes to the literature by providing empirical evidence of the role of social perception in hiring discrimination on a global scale. Using fictitious applications, however, may only partially capture the complexities of real-world hiring processes. Additionally, Veit et al. (2021) did not delve into the potential impact of other factors that could influ-

ence the results, such as the specific cultural context of each country or ascriptive characteristics and group affiliation of the job candidate. The manipulated statements in the cover letter are likely less salient and relevant to the hiring decision than other information. Last, the study did not include negative warmth and competence signals, limiting the detection of the effects of social perception.

Veit et al. (2021) study is relevant to the present study as it investigates the causal effects of social perception in hiring decisions. It also suggests the importance of ethnic stereotypes and group membership in hiring decisions. The present study will help fill the gap in research by focusing on the perceptions of warmth and competence of various categories, such as race, gender, sexuality, and disability.

Social perceptions also include stereotypes, and Pownall et al. (2022) examined the social stereotypes about pregnant women. Specifically, these stereotypes revolve around the idea that pregnant women have a "baby brain" that makes them less competent. The research questions focused on how these stereotypes affect the perception of pregnant women in society and are grounded in the Stereotype Content Model. The study included two surveys, with a total sample size of 644 participants, all UK residents. The variables include the perceived warmth and competence of pregnant women. The findings indicate that individuals perceive pregnant women as warm but not competent because of the so-called "baby brain" stereotype (Pownall et al., 2022). Pownall et al. (2022) concluded that the "baby brain" stereotype significantly impacts the perception of pregnant women's competence, suggesting that individuals assume that pregnancy negatively affects cognitive function; however, perceptions of warmth are comparatively high.

Pownall et al. (2022), however, did not fully explore the implications of these stereotypes for pregnant women in different contexts, such as the workplace. The researchers also used a homogenous sample of white British respondents, leaving a gap for future research to investigate how these stereotypes affect pregnant women's experiences in various settings. Pownall et al. (2022) also did not examine the potential intersectionality of these stereotypes with other factors such as race, age, or socioeconomic status. Pownall et al. (2022) study is relevant to our study as it provides insights into how social perception, specifically stereotypes, can affect the treatment of different groups, in this case, pregnant women. Our study aims to complement these findings by examining the impact of social perception on callback rates across various categories, including pregnancy.

In another study, Lippens et al. (2023) conducted a comprehensive meta-analysis of hiring discrimination based on recent correspondence experiments. The research questions revolved around the extent and nature of hiring discrimination across various social categories, including race, gender, age, religion, disability, and sexual orientation. The study filled a gap in the literature by providing a more nuanced understanding of hiring discrimination across different social categories and geographical regions. The researchers used a meta-analytic approach to synthesize the results of 169 correspondence studies conducted across all continents between 2005 and 2020. Lippens et al. (2023) studied the discrimination rates across different social categories and regions and found that hiring discrimination is most pronounced against ethnic minorities and older candidates, with moderate discrimination against women and religious minorities and relatively low discrimination against disabled and LGBTQ+ candidates.

Lippens et al. (2023) concluded that hiring discrimination is widespread, but its extent varies across different social categories and regions. The researchers also found that discrimination rates have remained relatively stable, suggesting that efforts to combat hiring discrimination have not succeeded. Lippens et al. (2023) acknowledged, however, that the methodological differences between the included studies may have influenced their findings. Much of the variability around the meta-analytic estimates remained unexplained. Moreover, Lippens et al. (2023) did not explore the underlying mechanisms of hiring discrimination, which could be a fruitful avenue for future research. The Lippens et al. (2023) study provides a broader context for our study, supporting our assumption that hiring discrimination is a pervasive issue that varies across different social categories. Our study aims to link these differential hiring effects with social perception. That is, instead of looking at each category separately, we aim to unify categories through a single measure.

## 4.6 Introduction to the Stereotype Content Model (SCM)

The Stereotype Content Model (SCM), proposed by Fiske et al., 2007, is a theoretical framework that seeks to understand how stereotypes, prejudice, and discrimination operate within society. The model suggests that stereotypes are not merely negative assumptions about outgroups but are instead complex social structures that encompass two primary dimensions: competence and warmth.

Competence is often associated with a group's perceived status, with high-status groups typically viewed as more competent. Warmth, on the other hand, is linked to the perceived level of competition a group presents, with groups seen as competitors typically viewed as less warm. This dynamic creates a space for mixed stereotypes, where some groups may be pitied for their low competence but high warmth, while others may be envied for their high competence but low warmth (Fiske et al., 2007).

The SCM is particularly relevant to our study as it provides a comprehensive analysis of racial discrimination in the hiring process. The SCM is flexible to underlying theories of discrimination, offering a model-agnostic approach to discrimination. This flexibility allows the SCM to capture various factors that drive discrimination, such as hostility toward foreign cultural norms and implicit discrimination.

The SCM also offers a valuable lens through which to examine the intersectionality of social identities. Intersectionality, a concept first coined by Crenshaw, 1989, refers to the interconnected nature of social categorizations such as race, class, and gender, which can create overlapping and interdependent systems of discrimination or disadvantage. In the context of the SCM, intersectionality can help us understand how multiple social identities interact to influence perceptions of warmth and competence.

For example, a study by Nicolas and Fiske, 2023, found that biased information integration occurs when people rate intersecting categories, with ratings being more similar to the constituent with more negative and extreme stereotypes. The study also found that emergent properties are more prevalent for novel targets and targets with incongruent constituent stereotypes. Emergent perceptions tend to be more negative and less about competence or sociability. This finding suggests that intersectionality can significantly influence the formation and content of stereotypes, which in turn can impact discrimination in the labor market.

Halper et al. (2019) investigated gender bias in caregiving professions and how perceived warmth plays a role in this bias. The study filled a gap in the literature on the underlying process behind adverse reactions towards men in caregiving professions. The study involved surveys with undergraduate and online participants to evaluate perceptions of men in caregiving occupations such as preschool teaching and social work. Participants rated the warmth, competence, likability, and hireability of men and women in these professions. Participants rated men as neither more nor less competent than women, but men were considered less warm, which mediated the relationship between gender and negative hireability responses (Halper et al., 2019).

Halper et al. (2019) concluded that stereotypes about men's warmth contribute to perceptions that men are less likable, suitable, and hireable for caregiving professions than their female counterparts. The study provides insights into the gender bias in caregiving professions and how perceived warmth plays a role in this bias and suggests that interventions are needed to increase men's participation in caregiving roles. The samples, however, were primarily comprised of Western participants in a college setting or through a low-quality online sample, limiting the generalizability of the findings. Further, Halper et al. (2019) used one name per gender, raising the question of whether the perceptions of warmth were related to gender or specifically to the chosen name. Finally, the study lacked behavioral measures, such as a firm actually calling back an applicant. Halper et al. (2019) recommend cross-cultural research to address generalizability and potential interventions for altering perceptions of men in caregiving fields.

The Halper et al. (2019) study provides insight into the impact of social perception on hireability. Along with Veit et al. (2021), the study suggests a significant mediating effect of warmth and competence between group membership and hireability outcomes. The findings hint at role congruity, suggesting individuals aim to match warm candidates with warm professions. Our study expands this scope by examining various professions and group identities.

Multiple-group membership can also influence perceptions of warmth and competence. Strinić et al. (2020) examined multiple-group membership and its impact on warmth and competence perceptions in the workplace. The research questions revolved around how the intersection of different social categories (e.g., gender, age, ethnicity) influences the perception of warmth and competence. Strinić et al. (2020) sought to fill a gap in the literature by examining the combined effect of multiple social categories on stereotype content. The study's central idea is that the perception of warmth and competence for individuals with multiple-group membership may be qualitatively different than that for single groups. Strinić et al. (2020) hypothesized that combining social categories can lead to unique stereotype content that is not merely the sum of the stereotypes associated with each category.

The Strinić et al. (2020) empirical study employed a 2 (gender: man, woman) $\times$ 2 (age: 30, 55) $\times$ 2 (ethnicity: Swedish, Arab) factorial design. The sample consisted of 133 job recruiters from Sweden, who rated the warmth and competence of individuals belonging to different combinations of social categories. The findings revealed that the combination of social categories could indeed lead to unique stereotype

content; for instance, older Arab women were perceived as warmer but less competent than younger Arab women (Strinić et al., 2020). Strinić et al. (2020) concluded that the perception of warmth and competence for individuals with multiple-group membership is qualitatively different than that for single groups. This finding is significant as it challenges the additive assumption of multiple categorization research and provides a more nuanced understanding of stereotype content.

One of the main weaknesses of the Strinić et al. (2020) study was the choice of stimuli, and the researchers acknowledged that they could not determine whether combining groups could move the specific combination to a different quadrant in the stereotype content model space. Strinić et al. (2020) recommended future researchers further explore the impact of multiple-group membership on stereotype content, using a broader range of social categories and more diverse samples. However, the Strinić et al. (2020) study provides a deeper understanding of how multiple-group membership influences social perception in the workplace. Our study includes multiple intersecting categories that are not always linearly related, which could affect our results and conclusions.

Researchers have also examined warmth and competence in relation to employment bias and homelessness. For example, Martinez et al. (2022) investigated employment biases in service contexts, mainly focusing on the perceived warmth and competence of individuals experiencing houselessness, the extent of these biases, and how they affect employability. The study filled a gap in the literature by exploring the intersection of houselessness and employment in the service industry, a relatively under-researched topic. The study included two groups: one of individuals with hotel management experience and another of a general sample of adults. Participants rated hypothetical job applicants' warmth, competence, and hireability, as perceived from resumes and audio recordings of social interactions. Martinez et al. (2022) found an effect of warmth on hireability moderated by gender and concluded that these biases significantly affect the employment opportunities of unhoused individuals, particularly in the service industry.

The Martinez et al. (2022) argued that these biases are deeply ingrained and are influenced by societal stereotypes about unhoused individuals. The study contributes to the literature by providing empirical evidence of these biases and their impact on employment opportunities for unhoused individuals. The study, however, was limited to homelessness in the hotel industry and may not generalize to other populations or contexts. The study is particularly relevant to our research as it provides

empirical evidence of the impact of social perception on employment outcomes. Further, the findings of Martinez et al. (2022) suggest that competence may only sometimes be the most crucial construct in hiring decisions, especially in customer service jobs.

Research has shown that the stereotype content model has validity across cultures. Strinić et al. (2021) investigated occupational stereotypes among a professional sample of recruiters and other employees on the two fundamental dimensions of the stereotype content model: warmth and competence. Strinić et al. (2021) surveyed professionals' ratings of preselected occupations and whether the two-dimensional warmth/competence space applies to occupational stereotypes in a European context, specifically Sweden. The study was unique as it includes prespecified common occupations for representativeness, unlike previous research where participants selected the included occupations (Halper et al., 2019; Veit et al., 2021). Participants rated warmth and competence attributes in preselected occupations. Factor and cluster analyses were employed to investigate the two-dimensional structure of the warmth/competence space and how and whether occupations cluster as predicted by the stereotype content model (SCM; Strinić et al. (2021)). The study included the largest and most common occupations, using the Swedish Standard Classification of Occupations, as a basis for selecting occupations.

The Strinić et al. (2021) found that almost all occupations had a clear two-factorial structure corresponding to the warmth and competence dimensions. A five-cluster solution appropriately depicted how occupations disperse on these dimensions. The study provides valuable insights into the treatment and preferences of various social groups and how such stereotypes might relate to hiring preferences (Strinić et al., 2021). However, the study was limited in the scope of its representativeness, as its participants included only about a hundred Swedish recruiters, although their finding lends empirical support to the universality of the stereotype content model, given that similar analyses have worked in the North American context. Further, future research should include implicit stereotypes to better model the selection process. That is, future studies should not rely on explicit signals of warmth and competence, but should attempt to manipulate signals based on their warmth and competence stereotypes that recruiters are expected to elicit before selecting candidates.

The Strinić et al. (2021) study is relevant to the present study as it provides an understanding of occupational stereotypes, which could influence how the stereotypes of job candidates can differentially affect their callback rates depending on the

industry or position for which they are applying. That is, recruiters could match the stereotype content of social groups and occupations. Understanding occupational stereotypes in a different context could provide a comparative perspective. Furthermore, the present study is also designed to examine the impact of social perception on callback rates, which aligns with the focus of Strinić et al. (2021) study on warmth and competence perceptions of occupations.

Employment status also influences perceived competence, and Okoroji et al. (2023) investigated the impact of employment status on the perceived competence of job applicants. The researchers hypothesize that unemployed individuals are perceived as less competent than their employed counterparts, affecting their hiring chances. The central idea is that being unemployed can perpetuate unemployment, suggesting a bias in hiring practices that could lead to overlooking potentially talented candidates (Okoroji et al., 2023). Okoroji et al. (2023) employed an experimental methodology where participants with hiring experience rated CVs of equally qualified candidates differing only in their employment status. The researchers measured the participants' willingness to interview and hire the candidates and their perceptions of the candidates' competence. The study has two parts: an initial study and a high-powered follow-up replicating the initial study in a different economic context characterized by increased job insecurity (Okoroji et al., 2023).

Okoroji et al. (2023) findings supported the hypotheses. The initial and follow-up studies showed that unemployed candidates are perceived as less competent than employed candidates, and this perceived lack of competence fully mediates the relationship between employment status and employment-related outcomes. Okoroji et al. (2023) suggested that the bias against unemployed candidates could lead to organizations missing out on talented individuals who would have been shortlisted if they were employed.

Okoroji et al. (2023) acknowledged, however, that their study design may only partially replicate typical recruitment scenarios where hiring managers and HR professionals may view dozens of CVs quickly. The decisions in the survey were only hypothetical, unlike an actual correspondence study. Further, their CVs did not include names, hiding gender, racial, and socioeconomic identities. These findings suggest that researchers should explore practical changes to CVs to reduce the perception of incompetence, such as removing dates from CVs and only including the duration of any employment alongside a description. The findings of the study are relevant to our research as they provide empirical evidence on the impact of

competence perceptions on callback rates in the labor market. We aim to fill the gap left by Okoroji et al. (2023) by analyzing the effects of warmth and competence in real hiring decisions, not just hypothetical ones.

In conclusion, the Stereotype Content Model offers a comprehensive and flexible framework for understanding discrimination in the labor market. By considering the dimensions of warmth and competence, as well as the intersectionality of social identities, the SCM can provide valuable insights into the complex dynamics of discrimination.

## 4.7 Materials and Methods

### Callback Data

One hundred ninety-one studies were gathered by combining those included in Lippens et al., 2023, and our screening process (Fig. C.1 shows a PRISMA diagram). For each study, we extracted information on the callback rates for each group, along with study-specific characteristics. Furthermore, we searched for published *raw* datasets for each study in the meta-analytic database. *Raw* means that data contain observations, including names or category signals and callback rates, for each resume sent in the experiments. We requested authors provide these raw data from their study for unpublished datasets. Table C.1 shows the datasets gathered. Additionally, for each category signal from the meta-analytic data and each name from the raw datasets, Prolific participants provided ratings of warmth and competence.

### Warmth and Competence Perceptions

Our research methodology involved the use of multiple surveys, which were designed to gather data on perceptions of warmth and competence associated with different social groups. The surveys were administered through the Prolific platform, a popular online platform for academic research.

For the category-based studies, participants saw the signal exactly as described in the study and rated each of them based on their perceived warmth and competence. Respondents were instructed to base their ratings on what they believed would be the average American's first impression of each individual. The survey acknowledged that these impressions could be formed based on limited information and might not accurately reflect the individuals' true warmth and competence. The survey for name-based studies followed a similar format but asked participants to rate names instead of specific lines from the resume. Again, participants were asked to provide

ratings based on what they thought the average American's first impression would be.

Participants rated names and categorical signals on 100-point scales of warmth and competence (1 = not at all, 100 = extremely), based on how the groups are viewed by American society. They read, "We are not interested in your personal beliefs, but in how you think they are viewed by others." As in all our studies, this instruction was intended to reduce social desirability concerns and to tap perceived cultural stereotypes.

**Sample Characteristics:** The Prolific participants constituted a non-convenience, compensated sample with a requisite North American cultural background. This criterion was pivotal, as shared cultural backgrounds are known to foster similar stereotype perceptions, ensuring the recruited sample's stereotypes aligned with those of the study's hiring decision-makers. Post-rating, participants were queried on demographics (race, gender, age) and quality controls (e.g., native language proficiency).

Specifically, for the sample rating names: 57.52% identified as female, with an average age of 37.62 years. The predominant ethnicity was White/Caucasian (62.38%), and the most common education level was a bachelor's degree (33.91%). A majority were in stable employment (52.1%), with the most reported income range being $25,000 to $49,999 (26.6%). A significant 97.5% were native speakers. For the sample rating categories: 50.1% identified as female, with 39.1% in the 25–34 age bracket. A larger majority were White/Caucasian (77.7%), and 31.2% held a bachelor's degree. Employment was stable for 50.5%, with office-focused roles accounting for 36.6%. The income range of $25,000 to $49,999 was most common (27.2%). A notable 77.2% identified as agnostic, atheist, or non-religious. Lack of resume review experience was reported by 66.3%, and a small fraction (4.95%) had current or past military service.

The choice to use a Prolific sample is grounded in literature suggesting that stereotypes are (i) influenced by cultural backgrounds and (ii) pervasively shared within a culture. Consequently, individuals with similar cultural backgrounds are likely to hold comparable stereotypes, regardless of their professional background. This assumption holds even when considering participants from varied professions, as our inquiry does not investigate industry-specific perceptions but rather aims to understand societal views on a particular social signal. While one might argue that

recruiters, owing to their training, could be less prone to stereotypical assessments in hiring contexts, the question we posed in the online survey transcends specific industries and focuses on broader societal perceptions. The specific wording was: “In your opinion, what does the average American think about this person? Even if you disagree. [signal, e.g. name]" Therefore, we do not anticipate significant variations in responses across samples drawn from different professional backgrounds.

Furthermore, leveraging stereotypes from one sample to predict behaviors in another offers a conservative approach to evaluating the impact of stereotypes on actions. This method likely leads to an underestimation of the effect size compared to directly measuring decision-makers' stereotypes. Our primary concern is with the influence of broad cultural stereotypes on decision-making processes. The extent to which individuals' actions reflect their personal stereotypes, which may not align with societal norms, represents a separate and potentially less consequential issue. This is because societal-level disparities arise when collective decision-making is guided by uniform assumptions based on shared stereotypes.

**Data Availability:** Raw data from published studies we used is publicly available, the authors included links to those files in their replication package. Furthermore, the authors of this study provide their raw data and analysis code on a public GitHub repository upon publication. Access to the currently private GitHub repository is granted upon request.

**Statistical analysis**

Analysis was carried out with R version 4.2.2. Meta models were estimated with packages `metafor_3.8-1` and `meta_6.2-1`. For meta-analytic analyses based on correlations, we deployed a random-effects model. Each correlation $r$ was transformed into Fisher's z: $z = 0.5 \log_e \left(\frac{1+r}{1-r}\right)$, to ensure that the sampling distribution was approximately normal. The model was adjusted via the Hartung-Knapp modification (Hartung and Knapp, 2001).

To estimate the random-effects model, the variance of the distribution of true effect sizes, $\tau^2$, had to be estimated, for which we deploy Maximum Likelihood (Viechtbauer, 2005). The confidence intervals around $\tau^2$ were estimated via the *Q-Profile* method (Veroniki et al., 2016). Furthermore, we calculated the $I^2$ statistic (Thompson and Higgins, 2002) to measure between-study heterogeneity.

Prediction intervals provide a valuable tool for estimating the likely range of effects that future studies may have based on the current evidence. As opposed to confidence intervals, prediction intervals consider $\tau^2$ to estimate the likely range of effects of future studies.

The meta-regressions 4.1 were specified as mixed-effects models: $\hat{\theta}_k = \theta + \beta x_k + \epsilon_k + \zeta_k$. The first error, $\epsilon_k$, represented the sampling error through which a study's effect size deviates from its true effect. The second error, $\zeta_k$, indicated that even the true effect size of a study is only sampled from an overarching distribution of effect sizes.

**Heterogeneity analysis**

We visualized the contribution of each study to the overall heterogeneity against its influence on the pooled effect size (Baujat plot, Fig. C.2). We also computed several influence diagnostics (Externally Standardized Residuals, DFFITS Value, Cook's Distance, Covariance Ratio, Leave-One-Out $\tau^2$, Hat Value, Study Weight, Fig. C.3). A leave-one-out robustness analysis was used to point to the study whose exclusion results in the largest decrease in the $I^2$ statistic (Fig. C.4). Additionally, we implemented a Graphical display of heterogeneity (GOSH) plot analysis (Fig. C.5).

**Intraclass correlation (ICC)**

We calculated the ICC through a two-way random-effects model (as provided by package `psych`) to assess the reliability of the average of $k$ ratings for each signal $i$. We described each rating as $y_{ij} = \mu + r_i + c_j + e_{ij}$, where $\mu$ was the average rating, $r_i \sim N(0, \sigma_r^2)$ and $c_j \sim N(0, \sigma_c^2)$ were random effects for the signals and raters, respectively, and $e_{ij}$ was the error term. Then, we computed ICC $= \frac{\sigma_r^2}{\sigma_r^2 + \left(\sigma_c^2 + \sigma_\epsilon^2\right)/k}$ (Liljequist et al., 2019).

**Finite mixture models (FMMs)**

We used an FMM to generate two latent classes with distinct effects of PC1 on callback rates, with the probability of belonging to class $i$ defined as $\pi_i = \frac{\exp(\gamma_i)}{\sum_{j=1}^{g} \exp\left(\gamma_j\right)}$, where $\gamma_i$ was a function of job characteristics (Table C.11).

## 4.8 Results

### Names: Social perception predicts callback in correspondence studies that vary names

We identified studies through a systematic search of correspondence experiments in North American labor markets (see PRISMA Flow Diagram, Fig. C.1). We further extracted name-specific callback rates from studies that reported or made them available through replication datasets for the following analyses. This procedure created a sample of eight studies.

Before examining warmth and competence, we first analyzed how callback varies by race and gender. The difference between groups was summarized by the ratio of the callback rates of the potentially discriminated-against group compared to the benchmark group, with a ratio of 1 indicating perfect parity, ratios $< 1$ indicating negative discrimination, and those $> 1$ indicating privileged treatment.

In our sample, the callback ratio was $\hat{\theta} = 0.798$ for Black names, which was significantly less than one ($p = .07$). The same ratio computed by Lippens et al., 2023, is $\hat{\theta} = 0.68$ ($p < .001$). For the female gender compared to male, our estimated ratios were $\hat{\theta} = 1.02$ ($p = .36$) in the eight studies we had, agreeing with Lippens' $\hat{\theta} = 1.02$ ($p < .003$) (Lippens et al., 2023). Together the data showed a $20-30\%$ reduction in callbacks for Black names and no reduction for female names (Table C.6). Our analysis did not differentiate between male and female-dominated occupations, which may account for the lack of a significant effect observed for females, as emphasized by Galos and Coppock, 2023.

To measure warmth and competence, lists of names from the correspondence studies were given to participants on Prolific (787 raters total, 85.9 per name). To evaluate the consistency of ratings across categories, we computed the intraclass correlation (ICC, as defined in Materials and Methods). Our results reveal that the level of agreement between raters differed across various studies, with agreement ranging from excellent to good in most studies (Table C.2). This variation was crucial, as low intraclass correlations of social categories create an upper bound on the reliability of the ratings (see Discussion for details).

The callback rates were computed by averaging the decisions of multiple hiring managers. Meanwhile, the warmth and competence scores were obtained from a different sample. To ensure reliable social perception measurements, we specifically recruited participants residing in North America with demographics closely resembling those of the average hiring manager, and we averaged ratings across raters.

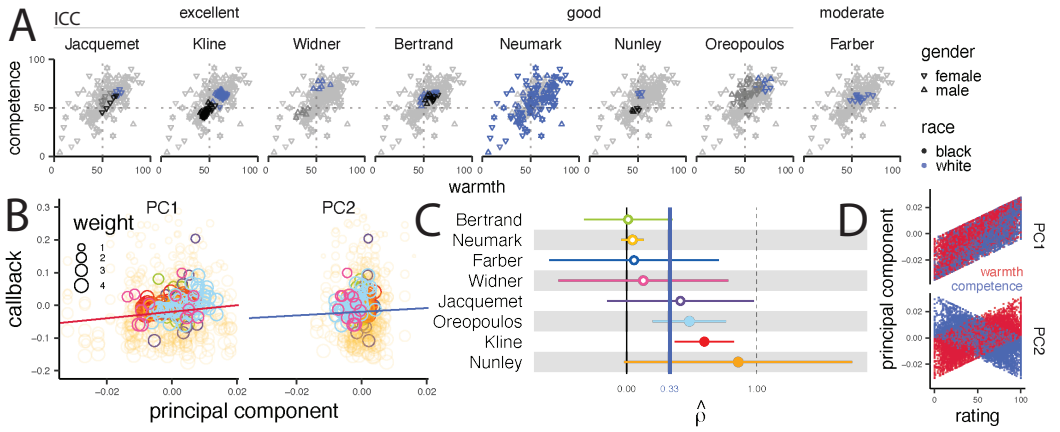



Figure 4.2: Warmth and competence ratings across names and their association with callback rates

(A) Each scatterplot shows warmth and competence for each name in the sample one study (with the first author name at the top). The correlations between the two rating scales are strongly positive in all eight studies (Table C.4). (B) Correlations between callback rates and PC1 and PC2 components associated with specific names (aggregating all studies). Data from different studies are identified by colors, with the legend shown in panel (C). The slope coefficients, shown in Table 4.1 are $\hat{\beta}_{\text{PC1}} = 1.00(p < .001)$, $\hat{\beta}_{\text{PC2}} = .56(p = .43)$. (C) Forest plot of confidence intervals for study-specific estimates of the correlation between callback rate and the first principal component PC1, $\hat{\rho}$(callback, PC1). All correlations are positive. The pooled effect is $\hat{\rho} = .33$. (D) Scatter plots of name-specific warmth and competence ratings showing the structure of PC1 and PC2.

This enabled us to confidently match the social perception ratings with the callback rates per name.

Those warmth and competence ratings, across names in different studies, are shown in Figure 4.2A. There were only minor differences in warmth or competence ratings between black and white candidates or males and females (between 2 and 7 points on the 100-point scale), except for a marginally significant difference in competence between black and white ($-11.52$, $p = .06$, Table C.6).

Figure 4.2A shows strong, reliable positive associations between warmth and competence within all eight studies, ranging from $.41 - .92$ (Table C.4). The pooled correlation is $\hat{\rho}_{w,c} = .78$ ($p < .001$). We, therefore, used principal component analysis (PCA) as a proxy for social perceptions. Figure 4.2D shows how the principal component scores (y-axis) are related to warmth and competence ratings (x-axis). The first component (PC1) reflects the positive association; it explains 79.3% of the variance. PC2 accounts for only 20.7% of the total variance, indicating its less

prominent role in the overall data structure. As a result, our subsequent analyses will focus on the PCs rather than the original ratings that generated them.

**Correlation between callback and PC1 as an effect** Figure 4.2C is a forest plot of the estimated correlations $\hat{\rho}$(callback, PC1) and 95% confidence intervals of the eight studies.

The effects across studies were pooled via a meta-analytic random effects model. The pooled correlation between the callback percentage and PC1 is $\hat{\rho} = .33$ ($p = .03$), indicating a moderate correlation (Cohen, 2013). To interpret the pooled effect size meaningfully, we must consider the variance of the true effect sizes distribution, $\tau^2$, and the between-study heterogeneity, $I^2$ (see Materials and Methods). As suggested by Figure 4.2C, there is "substantial heterogeneity" (Higgins and Thompson, 2002) among studies: 83 percent of the variation in effect sizes is due to between-study heterogeneity ($I^2 = .83$, 95% CI $[.68 - .91]$). Furthermore, the variance of the true effect sizes distribution is significantly greater than zero ($\tau^2 = 0.08$, 95% CI $[0.02 - 0.67]$).

Given the large level of heterogeneity in our analysis, we find a wide prediction interval (IntHout et al., 2016) (from $-.40$ to .81, details in Materials and Methods), suggesting that future studies could potentially reveal negative correlations. Therefore, caution is warranted in interpreting the results, and further research is needed to clarify the effect of social perception on callback.

In order to ensure the robustness of our findings and account for potential outliers, we conducted a comprehensive outlier and heterogeneity analysis. Only two out of eight tests (details in SI) identified outliers which, when excluded, re-estimate $\hat{\rho}$ as .22 or .34, both values remaining close to the .33 all-study estimate in Figure 4.2C. Furthermore, Egger's regression test (Fig. C.6; intercept = 1.8, 95% CI $[-0.25, 3.85]$, $t = 1.72$, $p = .14$) did not indicate bias.

As an alternative specification to the meta-analysis with $\hat{\rho}$, Table 4.1 reports results of a mixed effects model of callback rates against the PCs and raw ratings. The results are visualized in Figure 4.2B. The coefficient for PC1 is positive $\hat{\beta}_{\text{PC1}}$=1.00 and highly significant ($p = .0008$).

The correlations for warmth ($\hat{\rho} = .34$; $p = .02$) and competence ($\hat{\rho} = .26$, $p = .09$) are similar to those observed for the first principal component (PC1). This small warmth-competence difference is consistent with much evidence that judgments of

warmth are faster, more reliable, and more associated with behavior than competence judgments (e.g., Cuddy et al., 2007).

Table 4.1: Linear probability regressions of callback rates on principal components and social perception ratings

| | | 95% CI | | | |
|---|---|---|---|---|---|
| | **estimate** | **lower** | **upper** | **p-value** | **SE** |
| **Meta regression for categories**[1] | | | | | |
| intercept | -0.32 | -1.06 | 0.43 | 0.08 | 0.37 |
| PC1 | 1.16 | -0.28 | 2.59 | 0.12 | 0.72 |
| PC2 | -0.62 | -3.58 | 2.35 | 0.69 | 1.49 |
| **Meta regression for names**[2] | | | | | |
| intercept | -1.97 | -2.47 | -1.48 | 0.00 | 0.25 |
| PC1 | 1.00 | 0.41 | 1.58 | 0.00 | 0.30 |
| PC2 | 0.56 | -0.83 | 1.96 | 0.43 | 0.71 |
| **Correlations $\rho$(callback, variable) for names**[3] | | | | | |
| PC1 | 0.33 | 0.03 | 0.66 | 0.03 | 0.13 |
| warmth | 0.34 | 0.08 | 0.64 | 0.02 | 0.12 |
| comp | 0.26 | 0.06 | 0.58 | 0.09 | 0.14 |

**Note:** [1]Mixed-Effects Model (79 levels; $\tau^2$ estimator: ML)
[2]Mixed-Effects Model (691 names; $\tau^2$ estimator: ML)
[3]Three separate multivariate correlations; Random-Effects Model (8 studies; 725 observations); Inverse variance method, restricted maximum-likelihood estimator for $\tau^2$, Q-Profile method for the confidence interval of $\tau^2$ and $\tau$, Hartung-Knapp adjustment (df = 7), prediction interval based on t-distribution (df = 6), and Fisher's z transformation of correlations.

Moreover, we tested the predictive potential of our model for names. We found that $\hat{\rho}$(callback, PC1) in Jacquemet and Yannelis, 2012, is closest to the pooled effect, and we therefore used data from this study to make predictions. Specifically, we trained a linear model on all names except one and then used this model to predict the callback for the left-out name. Figure C.7 visually represents our findings, with lighter shades of blue indicating lower callback rates. Our analysis reveals that callback rates are highest in the upper quadrant of the warmth and competence scale, while the callback rates are lowest in the lowest quadrant of this plot. Interestingly, we also observe clusters of the race that the names would signal (Black, White, or foreign-sounding).

**Exploratory analysis points to differential effects of social perceptions across job types** Previous studies suggest a link between the stereotype content of occupations and group affliation (Cuddy et al., 2011; Halper et al., 2019). To investigate the role of social perceptions in determining callbacks across job types, we analyzed Farber et al., 2016, and Nunley et al., 2017, as those two studies provided adequate variation for meaningful conclusions. We employed finite mixture models (FMMs) to cluster occupations into two distinct groups for each study and examined the relationship between callbacks and PC1. Results reveal that the relationship between social perceptions and callbacks varies across job types. In Farber et al., 2016, the correlation between callbacks and PC1 was higher for service-oriented and less specialized jobs (r = .26 vs. r = .17). For Nunley et al., 2017, the correlations were slightly higher for advanced, specialized or managerial positions (r = .80 vs. r = .77).

An exploratory analysis using partial correlations indicated distinct relationships between warmth, competence, and callbacks for different job categories. In Farber et al., 2016, warmth was more strongly associated with callbacks in jobs with greater social interaction requirements ($r = .36$ vs. $r = .29$). Competence, conversely, exhibited a negative association with callbacks, more pronounced in professional or technical industries ($r = -.20$ vs. $r = -.16$). Nunley et al., 2017, revealed a more complex picture, with the relationship of warmth and callbacks appearing negative only for higher cognitive and technical skill jobs ($r = -.34$ vs. $r = .28$), while competence displayed a positive association, stronger in entry-level or sales-oriented roles ($r = .83$ vs. $r = .47$).

**Categories: Mixed effects of social perception on callback rates in correspondence studies manipulating social identity categories**

In the first section, we analyzed only correspondence studies that varied names. In this subsection, we look at correspondence studies varying other categories, such as religious affiliation or membership in the LGBTQ community (Fig. 4.1). Therefore, in the following analyses, we use extracted effect sizes.

Prolific participants (787 raters total, 99.1 by level) rated each signal on a scale from 0 to 100 within a category (e.g., how warm/competent they think a "treasurer in gay and lesbian alliance" would be, Figure 4.1, and 4.3A). The intraclass correlation (ICC) Bartko, 1966, values vary across categories. Only two categories scored *poor*, while the remaining scored either *moderate* or *excellent* (Fig. 4.3A). Note

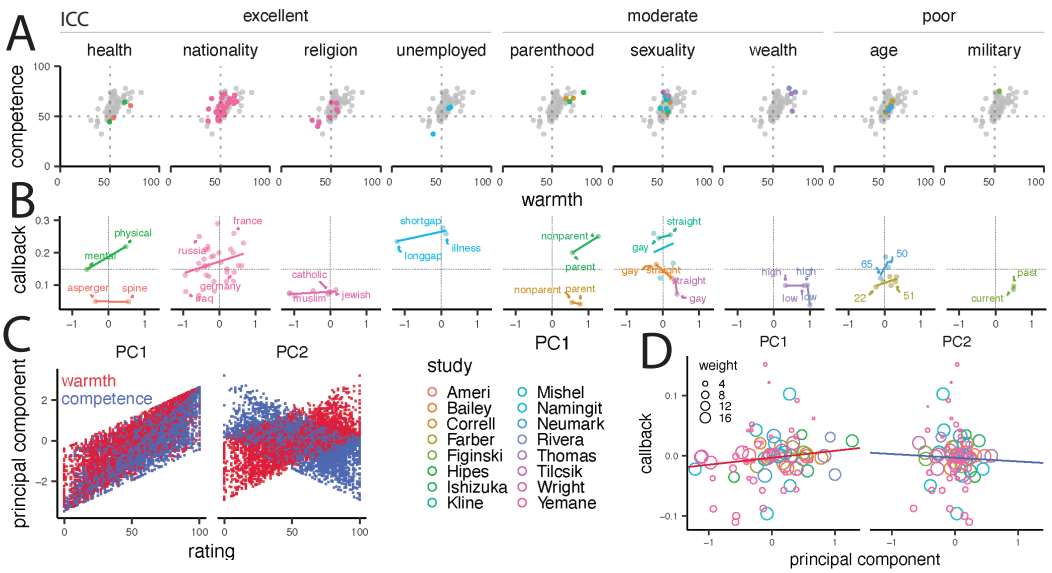



Figure 4.3: Warmth and competence ratings across categories and their association with callback rates

(A) Each scatterplot shows warmth and competence for each category-signal (with the category name at the top). The correlations between the two rating scales are strongly positive in all nine categories (Table C.5). (B) Linear regression of PC1 on callback by category and study. Data from different studies are identified by colors, with the legend shown in the center of row three. (C) Scatter plots of category-specific warmth and competence ratings showing the structure of PC1 and PC2. (D) Meta-regression of PC1 and PC2 on callback, where each circle identifies a signal by study; the circle size indicates the assigned weight in the meta-regression. Lines indicate fitted intercepts and $\hat{\beta}$s.

that, for sexuality and wealth, the different signals yielded vastly different ICCs, ranging from 0 to .83 (Table C.3). These findings suggest that raters' agreement levels varied across categories and signals.

Furthermore, we assessed the correlation between the warmth and competence ratings and found that the Pearson correlation index ($\rho$) was significant for most studies, with a pooled correlation of $\hat{\rho} = .595$ ($p < .001$, Table C.5). The correlation of target warmth and competence ratings across studies makes it challenging to estimate the independent effects of these dimensions of social perception on the callback rates. To better capture the variability in social perception of different social categories, we conducted a PCA, revealing two principle components.

PC1 explained 80.73% of the variability in warmth and competence ratings, combining the positively correlated measures onto a single dimension. PC2 represented negative associations and accounted for 19.27% of variance. Therefore, we analyze the principal components rather than the ratings which generated them (Fig. 4.3D).

In the following, we conduct a meta-regression pooling all studies to explore the extent to which the principal components predict callback. To increase the robustness of this analysis, we also perform a permutation test on our meta-regression models (Thompson and Higgins, 2002). The resulting estimate for the coefficient of the first principal component is .01, which is not statistically significant ($p$ = .134). Our model explains a small portion of the heterogeneity, accounting for only 3.31% (Table 4.1). Figure 4.3D visualizes the meta-regression model.

The meta-regression did not yield a significant overall effect. Therefore, we explored the relationship between PC1 and callback by category. Given the limited number of levels across categories (Fig. 4.3B), it was not possible to calculate a meaningful effect size directly relating ratings to callback at the study level. Thus, we present a graphical representation in Figure 4.3B, with lines representing fitted linear models for each study (Table C.7). For some categories, the relation between callback and PC1 is positive (e.g., nationality, which also samples a larger number—35—of categories). For most other categories, however, such as wealth, sexuality, and parenthood, there are both positive and negative slopes in different studies. Under the category of sexuality, slope signs in four studies were equally split between positive and negative, which is especially striking given the large range of ICCs across signals.

## 4.9 Discussion

Over the last few decades, many social scientists have used correspondence studies to document the disparities in outcomes that people experience in the labor market purely based on aspects of their social identity. Learning about these disparities is imperative for fostering fair and inclusive labor markets. Our study examined whether social perception predicts callback decisions in correspondence studies targeting the US and Canadian labor markets.

Jenkins et al., 2018 found that 12 distinct subdimensions of social perception (including friendliness, sincerity, self-control, efficiency, and others) could be effectively condensed into the two factors of warmth and competence (Cuddy et al., 2007, 2008). Building upon this finding, we focused our investigation on participants' perceptions of others' warmth and competence based on attributes offered in the relevant correspondence studies, such as name, religion, or sexual orientation. We found that the perceived warmth and competence of individuals' names or attributes were highly correlated, leading us to use the first principal component (PC1) to

measure favorable social perception. This component, reflecting both warmth and competence, explained about 70–81% of the variance in social perception ratings and was moderately linked to callback rates.

We found that in studies where names were varied to signal race, gender, and age, more favorable warmth and competence perception, based on names, positively predicted callback. However, for studies varying applicant characteristics such as sexuality and disability status, the effects of social perception on callback rates are ambiguous: some categories show a positive association between favorable social perception and callback rates, such as age and nationality, while others show a negative association. This result is unsurprising, given the small number of levels for some of the categories, and the effects observed for these categories are more susceptible to measurement issues like low inter-rater reliability (ICC).

The wide prediction interval for the positive correlation in our name analysis suggests that future studies might uncover negative correlations between positive ratings and callback rates. Our stringent selection criteria—restricting studies to those altering names to signify race and gender, conducted in North America, and offering raw data—resulted in a relatively small sample and excluded industry-specific variables. We found no publication bias in this sample. Moreover, our prediction approach adds variability, as it depends on perceptions from a group separate from the actual decision-makers, potentially contributing to the broad prediction interval.

Despite differences in the population that rated social perception and the employers making hiring decisions, there is a noteworthy predictive relationship between these ratings and callback rates, suggesting common cultural biases. The accuracy of these predictions could be even greater if the raters' demographics more closely matched those of the hiring decision-makers.

The reliability of categorical ratings in our study is measured by intraclass correlations (ICC). Certain social information categories, like military status and age, have shown low ICC, indicating raters' disagreement on the warmth and competence perceived in these groups. This disagreement suggests limitations in the predictive ability of social perception measurements for these categories. To enhance prediction, future studies could focus on gathering more ratings for categories with traditionally low ICC.

The validity of correspondence studies hinges on the subtle resume signaling of category membership being perceived by employers who read the resumes. Yet,

monitoring how much attention these signals receive is challenging outside of laboratory or carefully designed field settings, which can track attention more directly (Bartoš et al., 2016; Konovalov and Krajbich, 2020). Signal effectiveness likely varies by category, affecting study outcomes. This study aggregates results from various studies using different signals, indicating the large variance in signaling methods present in correspondence studies. Yet, future research is needed to determine the ways in which different signals modulate attention, and how such an effect may impact callback rates.

The discussion of discrimination theories in our study pertains to foundational economic models, delineating primarily into taste-based (Becker, 1971) and statistical discrimination (Arrow, 1998; Phelps, 1972) theories. Contrasting with these models, Bertrand and Mullainathan, 2004 suggest a “lexicographic search" pattern among hiring managers. In this approach, employers stop reading a resume once they encounter a salient signal (e.g., an African-American name). Our Prolific survey setup presents only the signals of interest (i.e., names or categories) for warmth and competence ratings. However, since certain signals are listed later on a resume (e.g., club membership), it is likely that the effect of warmth and competence perceptions generated by less salient signals may be overwhelmed by more salient ones (which may not always be associated with a discriminated group). This suggests that our model may overestimate the impact of less salient signals. Furthermore, the institutional discrimination theory posits that discrimination intensity is contextually determined (Brinton and Nee, 1998). Our study, however, is constrained to the North American job market context, and does not explore these contextual variations, presenting an opportunity for future research.

Our study’s practical implications lie in harnessing the link between social perceptions and callback rates to refine recruitment practices. Understanding discrimination via social perceptions facilitates generalization to underexplored stereotypes, crucial for protecting intersectional groups from bias. Perceptions associated with one group can inform on multiple intersecting identities (Nicolas and Fiske, 2023). Thus, our research advocates for a predictive model to anticipate labor market outcomes for intersectional groups, highlighting a direction for future bias mitigation efforts. Third, our study may contribute to the development of responsible AI by offering computer scientists insights into potential debiasing strategies proven effective in human decision-making, which can be translated into AI models. Most studies rely on sets of examples, e.g., various professions, to detect biases, thereby

lacking a validated collection for comprehensively assessing biases. In contrast, an approach grounded in social perception moves beyond sets of examples, providing a broader framework. A few authors in the representation learning literature have seen value in this approach (Otterbacher et al., 2017).

**Challenges in the Research Process**

The process of conducting this research and producing the paper "Social perceptions of warmth and competence predict experimental callback rates in North American labor market experiments" was marked by a series of challenges that required careful navigation and innovative solutions.

The first challenge encountered was the integration of data from a multitude of different studies. Each of these studies, with their unique methodologies, categories, and contexts, presented a distinct set of data. The task of harmonizing these disparate data sources into a unified dataset for analysis was a significant undertaking. It necessitated a profound understanding of each study's methodology and a meticulous approach to ensure the consistent treatment of data across all studies.

In tandem with this, the management of data from the 32 different studies posed logistical challenges. The unique data structures inherent to each study required the development and implementation of a robust data management system. This was particularly challenging given the imperative to maintain the integrity of the data from each study while simultaneously creating a unified dataset for the meta-analysis.

The extraction of data from the studies in a consistent and reliable manner was another hurdle that had to be overcome. The complexity of the process was such that it required multiple iterations before a satisfactory protocol for data extraction could be established. The challenge lay in developing a data extraction protocol that could be uniformly applied across all studies, despite their inherent differences.

A further complication arose from the need to separate name-based studies from category-based studies. The research encompassed studies that varied in their focus - some were based on names, signaling race and gender, while others were based on categories such as sexuality and disability. The task of segregating these different types of studies and analyzing them separately demanded careful consideration of the unique characteristics intrinsic to each type of study.

Another limitation of our study, and of correspondence studies in general, is the lack of information about the composition of the workforce at the companies that

received the resumes. The demographic makeup of existing employees could potentially influence hiring decisions and callback rates (e.g., a minority-owned business favoring minority applicants). Future research could benefit from incorporating this and other contextual variables.

Identifying the appropriate techniques for the analyses presented another challenge. The complexity of the data and the research questions necessitated the use of robust statistical techniques capable of handling the intricacies of the data and sensitive enough to detect the patterns of interest.

Beyond these methodological challenges, the research also demanded a deep understanding of the Stereotype Content Model and its application in the context of labor market outcomes. This required a thorough review of the literature and a careful application of the model to the data. Additionally, the ethical considerations associated with researching stereotypes and discrimination had to be navigated, ensuring that the research was conducted in a responsible and sensitive manner.

In summary, the process of conducting this research and producing this paper was marked by a series of challenges that required careful navigation, innovative solutions, and a deep understanding of both the data and the theoretical frameworks underpinning the research. The successful completion of this research, despite these challenges, is a testament to the rigorous and meticulous approach adopted by the research team.

## References


Agerström, J., Björklund, F., Carlsson, R., & Rooth, D.-O. (2012). Warm and competent hassan= cold and incompetent eric: A harsh equation of real-life hiring discrimination. *Basic and Applied Social Psychology*, *34*(4), 359–366.

Arrow, K. J. (1973). Higher education as a filter. *Journal of Public Economics*, *2*(3), 193–216. https://doi.org/10.1016/0047-2727(73)90013-3

Arrow, K. J. (1998). What has economics to say about racial discrimination? *Journal of Economic Perspectives*, *12*(2), 91–100. https://doi.org/10.1257/jep.12.2.91

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological reports*, *19*(1), 3–11.

Bartoš, V., Bauer, M., Chytilová, J., & Matějka, F. (2016). Attention discrimination: Theory and field experiments with monitoring information acquisition. *American Economic Review*, *106*(6), 1437–1475.

Becker, G. S. (1971, August 15). *The economics of discrimination* (2nd ed.). University of Chicago Press.

Bertrand, M., & Duflo, E. (2017). Field experiments on discrimination. *Handbook of Field Experiments*, *1*, 309–393. https://doi.org/10.1016/bs.hefe.2016.08.004

Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? a field experiment on labor market discrimination. *American Economic Review*, *94*(4), 991–1013. https://doi.org/10.1257/0002828042002561

Bessudnov, A., & Shcherbak, A. (2018, October 18). *Ethnic discrimination in multi-ethnic societies: Evidence from Russia* (preprint). SocArXiv. Retrieved June 26, 2023, from https://osf.io/2qzus

Bloch, K. R., Taylor, T., Church, J., & Buck, A. (2020). An intersectional approach to the glass ceiling: Gender, race and share of middle and senior management in U.S. workplaces. *Sex Roles*, *84*(5), 312–325. https://doi.org/10.1007/s11199-020-01168-4

Bogardus, E. S. (1925). Measuring social distance. *Journal of Applied Sociology*, *9*, 299–308. https://babel.hathitrust.org/cgi/pt?id=inu.30000104215342&view=1up&seq=303&q1=bogardus

Brinton, M. C., & Nee, V. (1998). Introduction. In *The new institutionalism in sociology* (pp. xv–xix). Russell Sage Foundation. https://www.jstor.org/stable/10.7758/9781610440837.5

Browne, I., & Misra, J. (2003). The intersection of gender and race in the labor market. *Annual Review of Sociology*, *29*(1), 487–513. https://doi.org/10.1146/annurev.soc.29.010202.100016

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic Press.

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, *1989*(1), 139–167. https://chicagounbound.uchicago.edu/uclf/vol1989/iss1/8

Cuddy, A. J., Fiske, S. T., & Glick, P. (2007). The bias map: Behaviors from intergroup affect and stereotypes. *Journal of personality and social psychology*, *92*(4), 631.

Cuddy, A. J., Fiske, S. T., & Glick, P. (2008, January 1). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. In *Advances in experimental social psychology* (pp. 61–149, Vol. 40). Elsevier. Retrieved May 30, 2024, from https://linkinghub.elsevier.com/retrieve/pii/S0065260107000020

Cuddy, A. J., Glick, P., & Beninger, A. (2011). The dynamics of warmth and competence judgments, and their outcomes in organizations. *Research in Organizational Behavior*, *31*, 73–98. https://doi.org/10.1016/j.riob.2011.10.004

Daniel, W. W. (1968). *Racial discrimination in England: Based on the PEP report* (Vol. 1084). Harmondsworth: Penguin.

Di Stasio, V., Lancee, B., Veit, S., & Yemane, R. (2019). Muslim by default or religious discrimination? Results from a cross-national field experiment on hiring discrimination. *Journal of Ethnic and Migration Studies*, *47*(6), 1305–1326. https://doi.org/10.1080/1369183X.2019.1622826

Farber, H. S., Silverman, D., & Von Wachter, T. (2016). Determinants of callbacks to job applications: An audit study. *American Economic Review*, *106*(5), 314–318.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Galos, D. R., & Coppock, A. (2023). Gender composition predicts gender bias: A meta-reanalysis of hiring discrimination audit experiments. *Science Advances*, *9*(18), eade7979. https://doi.org/10.1126/sciadv.ade7979

Hagendoorn, L. (1995). Intergroup biases in multiple group systems: The perception of ethnic hierarchies. *European Review of Social Psychology*, *6*(1), 199–228. https://doi.org/10.1080/14792779443000058

Hagendoorn, L., & Hraba, J. (1989). Foreign, different, deviant, seclusive and working class: Anchors to an ethnic hierarchy in the netherlands. *Ethnic and Racial Studies*, *12*(4), 441–468. https://doi.org/10.1080/01419870.1989.9993647

Halper, L. R., Cowgill, C. M., & Rios, K. (2019). Gender bias in caregiving professions: The role of perceived warmth. *Journal of Applied Social Psychology*, *49*(9), 549–562. https://doi.org/10.1111/jasp.12615

Hartung, J., & Knapp, G. (2001). A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine*, *20*(24), 3875–3889.

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

Hollis, L. P. (2018). Bullied out of position: Black women's complex intersectionality, workplace bullying, and resulting career disruption. *Journal of Black Sexuality and Relationships*, *4*(3), 73–89. https://doi.org/10.1353/bsr.2018.0004

IntHout, J., Ioannidis, J. P., Rovers, M. M., & Goeman, J. J. (2016). Plea for routinely presenting prediction intervals in meta-analysis. *BMJ open*, *6*(7), e010247.

Jacquemet, N., & Yannelis, C. (2012). Indiscriminate discrimination: A correspondence test for ethnic homophily in the Chicago labor market. *Labour Economics*, *19*(6), 824–832.

Jenkins, A. C., Karashchuk, P., Zhu, L., & Hsu, M. (2018). Predicting human behavior toward members of different social groups. *Proceedings of the National Academy of Sciences*, *115*(39), 9696–9701.

Konovalov, A., & Krajbich, I. (2020). Mouse tracking reveals structure knowledge in the absence of model-based choice. *Nature Communications*, *11*(1), 1893. https://doi.org/10.1038/s41467-020-15696-w

Lancee, B. (2019). Ethnic discrimination in hiring: Comparing groups across contexts. results from a cross-national field experiment. *Journal of Ethnic and Migration Studies*, *47*(6), 1181–1200. https://doi.org/10.1080/1369183X.2019.1622744

Liljequist, D., Elfving, B., & Skavberg Roaldsen, K. (2019). Intraclass correlation–A discussion and demonstration of basic features. *PloS one*, *14*(7), e0219854.

Lippens, L., Vermeiren, S., & Baert, S. (2023). The state of hiring discrimination: A meta-analysis of (almost) all recent correspondence experiments. *European Economic Review*, *151*, 104315. https://doi.org/10.1016/j.euroecorev.2022.104315

Martinez, L. R., Smith, N. A., Snoeyink, M. J., Noone, B. M., & Shockley, A. (2022). Unhoused and unhireable? Examining employment biases in service contexts related to perceived warmth and competence of people experiencing houselessness. *Journal of Community Psychology*, *50*(8), 3504–3524. https://doi.org/10.1002/jcop.22849

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*(1), 415–444. https://doi.org/10.1146/annurev.soc.27.1.415

Nicolas, G., & Fiske, S. T. (2023). Valence biases and emergence in the stereotype content of intersecting social categories. *Journal of Experimental Psychology: General*, *152*(9), 2520–2543. https://doi.org/10.1037/xge0001416

Nunley, J. M., Pugh, A., Romero, N., & Seals, R. A. (2017). The effects of unemployment and underemployment on employment opportunities: Results from a correspondence audit of the labor market for college graduates. *Ilr review*, *70*(3), 642–669.

Okoroji, C., Gleibs, I. H., & Howard, S. (2023). Inferring incompetence from employment status: An audit-like experiment. *PLOS ONE*, *18*(3), e0280596. https://doi.org/10.1371/journal.pone.0280596

Otterbacher, J., Bates, J., & Clough, P. (2017). Competent Men and Warm Women: Gender Stereotypes and Backlash in Image Search Results. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 6620–6631. https://doi.org/10.1145/3025453.3025727

Phelps, E. S. (1972). The statistical theory of racism and sexism. *The American Economic Review*, *62*(4), 659–661. Retrieved June 26, 2023, from https://www.jstor.org/stable/1806107

Pownall, M., Conner, M., & Hutter, R. R. C. (2022). Blame it on her 'baby brain'? investigating the contents of social stereotypes about pregnant women's warmth and competence. *British Journal of Social Psychology*, *62*(2), 692–707. https://doi.org/10.1111/bjso.12587

Quillian, L., & Lee, J. J. (2023). Trends in racial and ethnic discrimination in hiring in six western countries. *Proceedings of the National Academy of Sciences*, *120*(6), e2212875120. https://doi.org/10.1073/pnas.2212875120

Riach, P. A., & Rich, J. (2002). Field experiments of discrimination in the market place. *The Economic Journal*, *112*(483), F480–F518. https://doi.org/10.1111/1468-0297.00080

Rosette, A. S., Ponce de Leon, R., Koval, C. Z., & Harrison, D. A. (2018). Intersectionality: Connecting experiences of gender with race at work. *Research in Organizational Behavior*, *38*, 1–22. https://doi.org/10.1016/j.riob.2018.12.002

Strinić, A., Carlsson, M., & Agerström, J. (2020). Multiple-group membership: Warmth and competence perceptions in the workplace. *Journal of Business and Psychology*, *36*(5), 903–920. https://doi.org/10.1007/s10869-020-09713-4

Strinić, A., Carlsson, M., & Agerström, J. (2021). Occupational stereotypes: Professionals´ warmth and competence perceptions of occupations. *Personnel Review*, *51*(2), 603–619. https://doi.org/10.1108/PR-06-2020-0458

Tajfel, H. (1974). Social identity and intergroup behaviour. *Social Science Information*, *13*(2), 65–93. https://doi.org/10.1177/053901847401300204

Thatcher, S. M. B., Hymer, C. B., & Arwine, R. P. (2023). Pushing back against power: Using a multilevel power lens to understand intersectionality in the workplace. *Academy of Management Annals*, *17*(2), 710–750. https://doi.org/10.5465/annals.2021.0210

Thompson, S. G., & Higgins, J. P. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine*, *21*(11), 1559–1573.

Veit, S., Arnu, H., Di Stasio, V., Yemane, R., & Coenders, M. (2021). The "big two" in hiring discrimination: Evidence from a cross-national field experiment. *Personality and Social Psychology Bulletin*, *48*(2), 167–182. https://doi.org/10.1177/0146167220982900

Veit, S., & Thijsen, L. (2021). Almost identical but still treated differently: Hiring discrimination against foreign-born and domestic-born minorities. *Journal of Ethnic and Migration Studies*, *47*(6), 1285–1304.

Verkuyten, M., Hagendoorn, L., & Masson, K. (1996). The ethnic hierarchy among majority and minority youth in the netherlands. *Journal of Applied Social Psychology*, *26*(12), 1104–1118. https://doi.org/10.1111/j.1559-1816.1996.tb01127.x

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., Langan, D., & Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research synthesis methods*, *7*(1), 55–79.

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, *30*(3), 261–293.

*Chapter 5*

# CONCLUSION

This dissertation has described the research that explored the potential of integrating cognitive and behavioral science insights into policy-relevant domains, including labor market discrimination and math education. The studies sought to answer the following research questions: (1) In what ways can the principles of reinforcement learning, derived from computational psychology, be applied to real-world policy designs, including educational interventions? (2) How can insights into learning processes inform the development and improvement of digital education platforms and policies? (3) How can insights into cognition and behavior enhance the design and implementation of policies to reduce discrimination?

This dissertation presented the research in three empirical chapters, each addressing a specific aspect of psychologically informed policy research. First, the Introduction provided a brief background for this work. Chapter 2 applied computational models, including Q-learning, to field data from a large online math teaching platform, providing insights into teachers' adaptive decision-making processes. Chapter 3 analyzed field data from the same platform, using the resulting insights to design and evaluate behavioral interventions to improve student learning outcomes. Chapter 4 investigated the effects of perceived warmth and competence on callback rates in labor market field experiments, offering a novel perspective on the mechanisms underlying discrimination.

While each chapter focused on a distinct research question, methodology, and policy domain, they are interconnected by the overarching theme of leveraging cognitive and behavioral sciences to inform policy-making. Each chapter presented unique findings derived from its particular context and innovative approach, and further details are provided below.

## 5.1 Chapter 2: Reinforcement Learning Models in Teachers' Decision-Making

Chapter 2 introduced a new application of reinforcement learning (RL) models to gain insight into teachers' decision-making processes on the Zearn online math-teaching platform. By fitting Q-learning and actor-critic models to field data, this research demonstrated the potential of RL to capture the complex, adaptive nature of

teachers' pedagogical strategies. The RL approach was advantageous compared to traditional methods, as it allowed for a more systematic understanding of teachers' adaptive decision-making in a complex educational environment. The findings revealed distinct patterns of teaching behavior and highlighted the importance of individual differences in learning rates and decision-making processes for student outcomes.

## 5.2 Chapter 3: Data-Driven Behavioral Interventions in Math Education

Chapter 3 combined an unsupervised learning technique (i.e., Independent Component Analysis, ICA) and a large-scale field experiment to generate effective behavioral interventions for teachers on the Zearn online math-teaching platform. The Independent Component Analysis (ICA) identified key dimensions of teacher behavior that informed the design of targeted "empathy" and "habitization" interventions. The resulting interventions demonstrated the effectiveness of targeted behavioral interventions in enhancing student engagement and performance. The results showed significant improvements in student lesson completion and reduced learning struggles.

## 5.3 Chapter 4: Perceived Warmth and Competence in Labor Market Discrimination

Chapter 4 used the Stereotype Content Model to explain the differences in callback rates across groups in North American labor market experiments. Using two psychological measures (i.e., warmth and competence) helped to explain the varied impact of social perception on hiring discrimination, offering a novel perspective on the mechanisms underlying discrimination. Specifically, perceived warmth and competence explained a substantial portion of the variation in callback rates across different demographic groups.

## 5.4 Significance of the Studies

It is expected that the research findings presented in these chapters will contribute to a more comprehensive understanding of how cognitive and behavioral sciences can be integrated into policy-relevant domains to address complex societal challenges. By applying innovative methodologies and computational models to real-world data, this work demonstrated the feasibility and value of bridging the gap between these fields and policy-making. Across the chapters, several overarching themes emerged, including the importance of data-driven approaches, interdisciplinary collaboration,

and behavioral analysis to drive positive change through informed interventions. This is significant because these themes showcase the expansive scope and potential of research and innovation at the intersection of cognition, behavior, and public policy. This work's concepts and methods can apply to real-world policy issues beyond labor discrimination and education.

Furthermore, this research contributes to the growing toolkit of psychologically informed policy research. For example, paradigms traditionally associated with lab experiments (i.e., fitting animal behavior to a reinforcement learning model) worked well in a real-world context.

## 5.5 Limitations of the Studies

While the studies presented in these chapters contribute significantly to their respective fields, they also reveal limitations. For example, Chapter 4 focuses primarily on the North American context, which may limit the generalizability of the findings to other cultural and societal contexts. Similarly, the sample size and exclusion criteria used in Chapters 2 and 3 may restrict the representativeness of the results.

In addition to the topic-specific limitations, there are also some overarching limitations to the research approach and methodology used in the studies. One hurdle is the complexity and variability inherent in neural and behavioral data. Pollak and Wolfe (2020) note that the multifaceted factors that influence brain development and function, including genetics, environment, and social aspects, make it challenging to derive simple, one-size-fits-all policy solutions. Additionally, the dynamic nature of the brain, with critical periods of development and plasticity, adds another layer of complexity. Pollack and Wolfe also stress that high-quality studies require significant technical and infrastructural requirements and sustained funding, limiting accessibility for some researchers and policymakers. Furthermore, transparent communication about the limitations and uncertainties of neuroscientific research is crucial to avoid oversimplification or misinterpretation of findings, which could lead to misguided policies or unrealistic expectations (Farah, 2018).

## 5.6 Recommendations for Future Research, Practice and Policy

Nevertheless, the findings and limitations evident in each chapter point to several promising avenues for future research in psychologically informed policy-making. For example, future studies could explore the applicability of the Stereotype Content Model to other forms of discrimination, such as housing or education, and investigate the effectiveness of interventions designed to mitigate the impact of so-

cial perceptions on decision-making in these contexts. Furthermore, future studies might explore new policy domains and refine methodological approaches. For instance, more sophisticated computational models may incorporate a broader range of behavioral and contextual factors. Also, longitudinal studies may better assess the long-term impacts of interventions on individual and societal outcomes. Other research directions include the development of scalable, evidence-based interventions that can be widely implemented to improve public health, education, and social welfare. Further, integrating neuroscience and related technologies, such as wearable devices, offers exciting opportunities for better understanding cognitive function, tailoring interventions to individual needs, and gathering physiological data to inform policy decisions.

Nevertheless, the findings from this research have generated interesting policy implications and recommendations. First, based on the findings on the role of perceived warmth and competence in predicting callback rates, policymakers may consider implementing training programs that raise awareness about the impact of social perceptions on hiring decisions and promote more objective evaluation processes. These results may also contribute to developing responsible, unbiased AI. Notably, this approach provides a broader framework to analyze and compare perceptions across diverse groups, especially within under-researched, intersectionally stereotyped groups.

Moreover, based on the results from the RL models and data-driven interventions in Chapters 2 and 3, policymakers and educators might consider incorporating data-driven insights and targeted behavioral interventions into their pedagogy, such as those focused on empathy and habit formation, to enhance student engagement and performance. This effort could involve the development of adaptive learning systems that tailor content and support to individual students' needs, as well as implementing professional development programs that promote effective teaching strategies and data-driven decision-making.

As evidenced by this study, fostering interdisciplinary collaboration among researchers, policymakers, and practitioners and prioritizing data-driven approaches in policy design and implementation is crucial to fully realizing the potential of psychologically informed policy-making. This effort requires the development of shared frameworks, methodologies, and communication strategies that facilitate the exchange of knowledge and expertise across disciplinary boundaries. As psychologically informed policy research continues to evolve, it will be essential to maintain

a strong focus on interdisciplinary collaboration and actively seek out opportunities for innovation and knowledge exchange.

## 5.7 Summary

In conclusion, these studies have sought to contribute significantly to understanding how cognitive and behavioral sciences can be integrated into policy-relevant domains to address pressing societal challenges. By applying innovative methodologies and computational models to real-world data, this work has demonstrated the feasibility and value of bridging the gap between these fields and policy-making. The key findings from this work include the importance of data-driven approaches, the value of interdisciplinary collaboration, and the potential of behavioral interventions to drive positive change in various policy contexts. These insights provide a foundation for continued research and innovation at the intersection of cognition, behavior, and public policy and highlight the need for more evidence-based solutions to complex societal issues. The insights and approaches presented in this dissertation provide a foundation for continued research and innovation in psychologically informed policy-making. As our understanding of the brain continues to evolve, so will the opportunities for leveraging this knowledge to create more effective, evidence-based policies that promote the well-being of individuals and communities worldwide.

## References

Farah, M. J. (2018). Socioeconomic status and the brain: Prospects for neuroscience-informed policy. *Nature Reviews Neuroscience*, *19*(7), 428–438. https://doi.org/10.1038/s41583-018-0023-2

Pollak, S. D., & Wolfe, B. L. (2020). How developmental neuroscience can help address the problem of child poverty. *Development and Psychopathology*, *32*(5), 1640–1656. https://doi.org/10.1017/S0954579420001145

*Appendix A*

# APPENDIX FOR CHAPTER 2

## A.1 Supplementary Methods

### PCA vs. NMF

Principal Component Analysis (PCA) was our first methodological choice. It is widely utilized but assumes data normality (Jolliffe & Cadima, 2016) and maximizes variance explained, potentially overlooking subtle relationships between variables. Consequently, we also employed NMF, which, by contrast, imposes a non-negativity constraint and is more closely related to clustering algorithms, creating a more interpretable, sparse representation of behaviors (Ding et al., 2005; Lee & Seung, 1999). This technique is particularly advantageous for data representing counts or frequencies. By trying different techniques, we can explore the reduced-dimension representation best suited to our specific dataset and research questions.

### Temporal Dynamics

Our investigation into temporal dynamics confirmed the impact of lagged rewards and actions on decision-making: shaping future decisions by past experiences. Figure A.9 illustrates this relationship, showcasing the predictive accuracy and model fit across fixed-effects models with different lags, with BIC and AUC scores for the models with one-week lags as the baseline. The results suggest a preference for a lag of two periods as optimal, based on the "elbow" in the AUC curves and the minima in the BIC curves.

### Correlations Between Variables

We began to unveil the intricate relationships among the variables under consideration through a comprehensive correlation analysis, as depicted in Figure A.1. This correlation matrix elucidated the magnitude and direction of associations among variables such as badges earned, minutes spent per student, tower alerts, the number of students, and teacher minutes. These interconnections informed the construction of our reinforcement learning models by suggesting the influence of teacher effort on student achievement. In this correlation matrix, each cell represents the Spearman correlation coefficient between a pair of variables. The color and size of the circles in each cell reflect the strength and direction of the correlation, with blue indicating

positive correlations and red indicating negative correlations. The histograms along the diagonal provide a visual representation of the distribution of each variable.

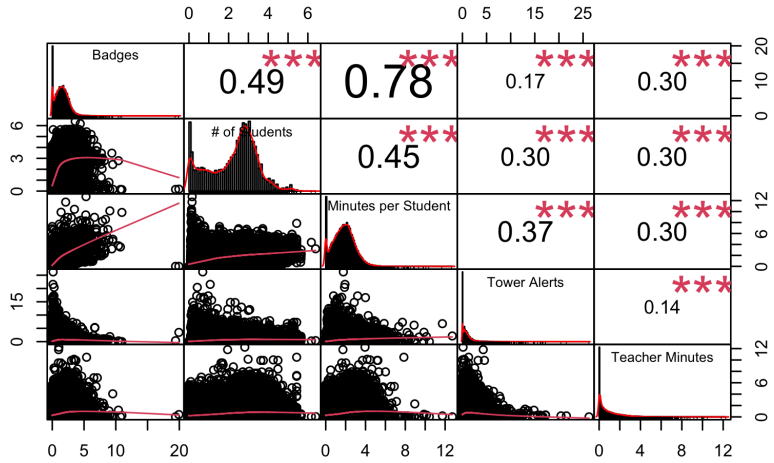



Figure A.1: Correlation coefficients between variables after standardization

## A.2 Supplementary Tables

### Teacher Variables

Table A.1: Catalog of teacher activities.

This table presents teachers' actions, including curriculum engagement, downloads of pedagogical materials, and completion of various interactive components within the Zearn educational platform.

| Variable | Description |
|---|---|
| PD Course Guide Download (Zearn, 2023, 2024k) | Detailed agenda for Professional Development (PD) courses focusing on classroom implementation, leadership, supporting diverse learners, using data to inform teaching practices, and accelerating student learning. |
| PD Course Notes Download (Zearn, 2023, 2024k) | Professional development session notes offering insights into effectively using Zearn's curriculum. |
| Curriculum Map Download (Zearn, 2024d) | Detailed outline of learning objectives and content. Presents a sequence of interconnected math concepts across grades, aligning with states' instructional requirements. |
| Assessments Download (Zearn, 2024b) | Assessments to evaluate student understanding of the material, including ongoing formative assessments, digital daily checks, and paper-based unit assessments. |
| Assessments Answer Key Download (Zearn, 2024l) | Solutions for assessments to aid in grading and feedback. Provides detailed rubrics for mission-level assessments. |
| Elementary Schedule Download (Zearn, 2024s) | A recommended schedule for elementary school-level Zearn curriculum activities to guide daily and weekly instructional planning, ensuring comprehensive coverage of curriculum content. |
| Grade Level Overview Download (Zearn, 2024n) | Provides a summary of learning objectives, pacing guidance, key grade-level terminology, a list of required materials, and details on the standards covered by each lesson. |
| Kindergarten Schedule Download (Zearn, 2024r) | Recommended schedules for Kindergarten, supporting structured instruction planning. |

*Continued on next page*

Table A.1 (continued)

| Variable | Description |
|---|---|
| Kindergarten Mission Download (Zearn, 2024h) | Details interactive activities focused on kindergarten-level concepts and their learning objectives. |
| Mission Overview Download (Zearn, 2024n) | Outlines a mission's flow of topics, lessons, and assessments; highlights foundational concepts introduced earlier; lists recently introduced terms and required materials for teacher-led instruction. |
| Optional Homework Download (Zearn, 2024q) | Assignments for additional practice, enhancing student learning outside of class. |
| Optional Problem Sets Download (Zearn, 2024m) | Exercises for extra practice, tailored to reinforce lesson concepts. |
| Small Group Lesson Download (Zearn, 2024i) | Lessons designed for small-group engagement. |
| Student Notes and Exit Tickets Download (Zearn, 2024o, 2024t) | Student notes supplement digital lessons with paper-and-pencil activities. Exit tickets are lesson-level assessments for teachers to monitor daily learning. |
| Teaching and Learning Approach Download (Zearn, 2024p) | Resources outlining Zearn's pedagogical methods. |
| Whole Group Fluency Download (Zearn, 2024j) | Lesson-aligned practice activities to build math fluency through whole-class engagement. |
| Whole Group Word Problems Download (Zearn, 2024i) | Word problem-solving activities intended for collaborative, whole-class engagement. |
| Fluency Completed (Zearn, 2024j) | Indicates teacher completed a fluency activity, typically given to students before their daily digital lessons. |
| Guided Practice Completed (Zearn, 2024e) | Indicates teacher completed a guided practice segment, where students learn new concepts. These include videos with on-screen teachers, interactive activities, and paper-and-pencil Student Notes. |

*Continued on next page*

Table A.1 (continued)

| Variable | Description |
|---|---|
| Kindergarten Activity Completed (Zearn, 2024g) | Indicates teacher completed an activity within the Kindergarten curriculum. |
| Number Gym Activity Completed (Zearn, 2024a) | Indicates teacher completed a Number Gym, an individually adaptive activity that builds number sense, reinforces previously learned skills, and addresses areas of unfinished learning. |
| Tower Completed (Zearn, 2024f) | Indicates teacher completed a Tower of Power, an activity that requires full mastery of lesson objectives and that students must complete independently. |
| Tower Struggled (Zearn, 2024c) | Indicates teacher made a mistake when engaging with the Tower of Power activity in a student role, triggering a "boost" (scaffolding remediation). |
| Tower Stage Failed (Zearn, 2024f) | Indicates teacher received three consecutive "boosts" due to repeated errors when engaging with the Tower of Power in a student role. |

Table A.2: Relationship between RL Parameters and Classroom Characteristics

The table presents the results of five regression models examining how reinforcement learning (RL) parameters predict income and poverty levels, number of students, number of classes per teacher, and whether the school had a paid Zearn subscription. RL parameters were estimated through non-hierarchical maximum likelihood estimation. Coefficients and standard errors (in parentheses) are provided for each parameter. Income and Poverty are treated as ordinal variables, while Total Students and Number of Classes are count variables. Paid Account is a binary variable. All RL parameters are standardized (z-scored) before analysis.

| | *Dependent variable:* | | | | |
|---|---|---|---|---|---|
| | Income | Poverty | Total Students | No. of Classes | Paid Account |
| | *ordered logistic* | *ordered logistic* | *Poisson* | *Poisson* | *logistic* |
| | (1) | (2) | (3) | (4) | (5) |
| Learning Rate ($\alpha$) | −0.119* | 0.032 | 0.006 | 0.021 | 0.208* |
| | (0.049) | (0.056) | (0.006) | (0.019) | (0.084) |
| Discount Factor ($\gamma$) | 0.021 | 0.112 | −0.006 | −0.038 | 0.003 |
| | (0.057) | (0.064) | (0.007) | (0.022) | (0.104) |
| Inverse Temperature ($\tau$) | −0.018 | −0.066 | 0.007 | −0.023 | 0.093 |
| | (0.050) | (0.053) | (0.006) | (0.021) | (0.080) |
| Initial Q-value | 0.232*** | 0.086 | 0.001 | −0.007 | −0.239* |
| | (0.056) | (0.063) | (0.007) | (0.021) | (0.099) |
| Cost | −0.260*** | 0.218*** | −0.014* | −0.059** | 0.609*** |
| | (0.052) | (0.061) | (0.006) | (0.020) | (0.122) |
| Constant | | | 3.027*** | 0.756*** | 2.139*** |
| | | | (0.005) | (0.016) | (0.086) |
| Observations | 1,737 | 1,668 | 1,782 | 1,782 | 1,782 |
| Log Likelihood | | | -5,728.112 | -2,671.142 | -628.732 |
| Akaike Inf. Crit. | | | 11,468.220 | 5,354.284 | 1,269.464 |
| *Note:* | | | | | *p<0.05; **p<0.01; ***p<0.001 |

| | Dependent Variable: Tower Alerts | | |
|---|---|---|---|
| **Parameter** | **(1)** | **(2)** | **(3)** |
| $\alpha$ | 0.061*** | 0.059*** | 0.050*** |
| | (0.015) | (0.015) | (0.015) |
| $\gamma$ | 0.006 | 0.015 | 0.023 |
| | (0.017) | (0.017) | (0.017) |
| $\tau$ | -0.019 | -0.023 | -0.023 |
| | (0.015) | (0.016) | (0.015) |
| Cost | 0.031* | 0.031 | 0.042* |
| | (0.016) | (0.021) | (0.021) |
| Starting Q-value | -0.019 | -0.013 | -0.002 |
| | (0.017) | (0.019) | (0.018) |
| No. of Weeks | | 0.008** | 0.007** |
| | | (0.003) | (0.003) |
| No. of Students | | -0.006** | -0.008*** |
| | | (0.002) | (0.002) |
| No. of Classes | | 0.066*** | -0.007 |
| | | (0.013) | (0.016) |
| Charter School | | | -0.026 |
| | | | (0.052) |
| Paid Zearn Account | | | 0.130*** |
| | | | (0.038) |
| Constant | 0.934*** | 0.730*** | -0.157 |
| | (0.013) | (0.080) | (0.172) |
| Control for AIC | | Yes | Yes |
| Control for Grade Level | | | Yes |
| Control for Poverty Level | | | Yes |
| Observations | 1,782 | 1,782 | 1,668 |
| $R^2$ | 0.026 | 0.051 | 0.162 |
| Adjusted $R^2$ | 0.023 | 0.046 | 0.153 |
| RSE (df) | 0.534 (1776) | 0.527 (1772) | 0.498 (1649) |
| F Statistic (df) | 9.479*** (5; 1776) | 10.646*** (9; 1772) | 17.683*** (18; 1649) |

*Note:* *p<0.05; **p<0.01; ***p<0.001

Table A.3: Impact of Q-learning model parameters on average weekly tower alerts per tower completion.

Three linear regression models examine the correlations between a teacher's reinforcement learning (RL) parameters and student struggle, measured by average weekly Tower Alerts. RL parameters were estimated through non-hierarchical maximum likelihood estimation. Model 1 includes only RL parameters. Model 2 adds controls for AIC, number of weeks, total students, and number of classes. Model 3 further incorporates controls for grade level, poverty level, charter school status, and whether the school has a paid Zearn account. Coefficients and standard errors (in parentheses) are provided for each parameter. RSE = Residual Standard Error.

| Parameter | Statistic | Non-hierarchical | Hierarchical | |
|---|---|---|---|---|
| | | | Individual | Hyperparameter |
| $\alpha$ | Mean | 0.353 | 0.0550 | 0.0549 |
| | Median | 0.423 | 0.0478 | |
| | 95% CI | [0.339, 0.367] | [0.0527, 0.0575] | [0.0364, 0.0767] |
| $\gamma$ | Mean | 0.448 | 0.0590 | 0.0582 |
| | Median | 0.462 | 0.0499 | |
| | 95% CI | [0.442, 0.453] | [0.0568, 0.0612] | [0.0273, 0.1114] |
| $\tau$ | Mean | 2.103 | 32.836 | 32.8356 |
| | Median | 1.752 | 32.8304 | |
| | 95% CI | [2.009, 2.202] | [32.8340, 32.8381] | [23.6889, 46.4811] |
| Cost | Mean | 0.596 | 0.0131 | 0.0131 |
| | Median | 0.638 | 0.0134 | |
| | 95% CI | [0.571, 0.622] | [0.0130, 0.0132] | [0.0082, 0.0198] |
| Initial Q-value | Mean | 0.651 | 0.0000 | 0.000 |
| | Median | 1.021 | -0.0037 | |
| | 95% CI | [0.605, 0.696] | [-0.0018, 0.0017] | [-0.0080, 0.0068] |

Table A.4: Comparison of Q-learning model parameters across estimation methods.

The table presents parameter estimates from non-hierarchical maximum likelihood and hierarchical Bayesian methods. Non-hierarchical estimates represent parameters fitted independently for each teacher, where CI is the confidence interval around the mean. For the hierarchical model, the individual column shows the means and medians of individual-level parameters with the confidence intervals around each mean. The hyperparameter column shows the estimated population-level means and their 95% credible intervals. All parameters are reported in their transformed space (logit transformation for $\alpha$ and $\gamma$, log transformation for $\tau$ and cost).

## A.3 Supplementary Figures

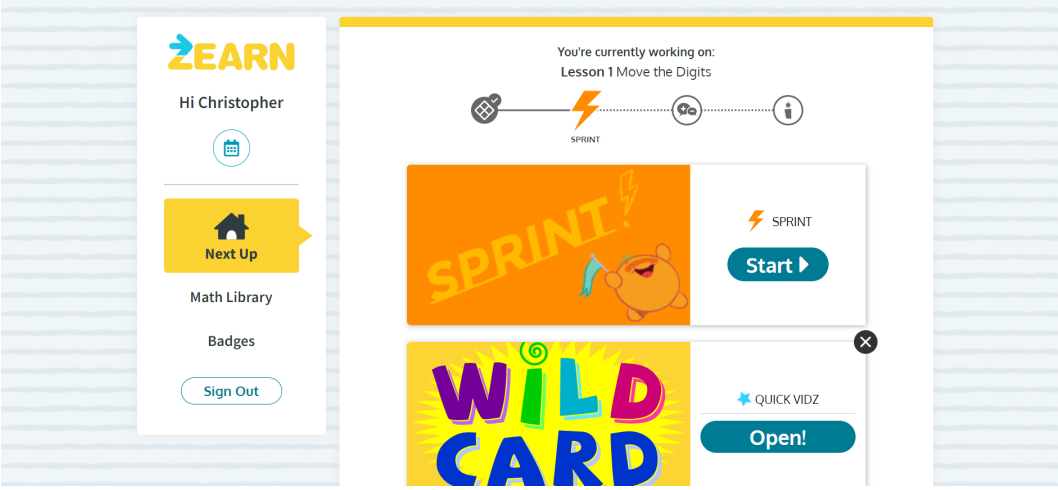



Figure A.2: Zearn student portal

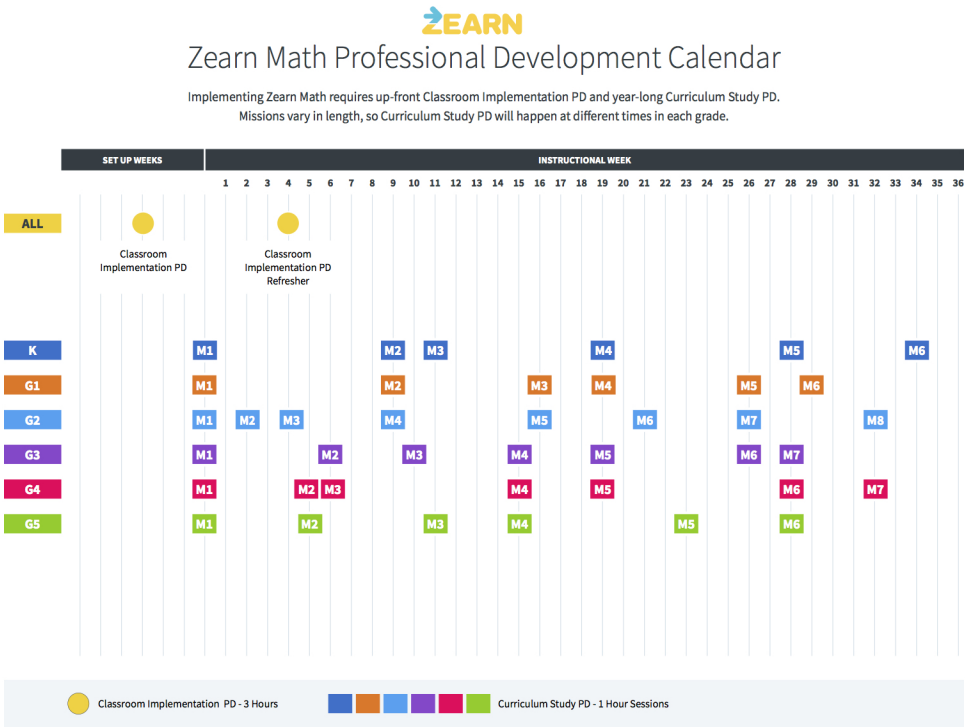
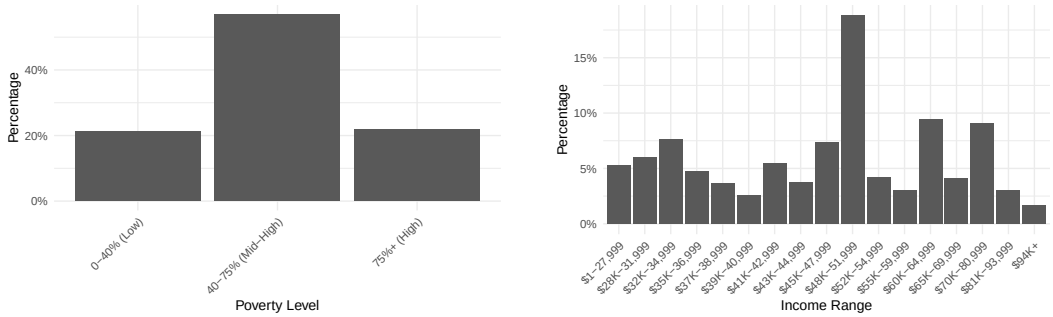



Figure A.3: Professional development calendar




(a) School poverty distribution (b) Median income distribution

Figure A.4: Distributions of school socioeconomic profiles

The first graph (a) categorizes schools into three groups based on the percentage of students eligible for free or reduced-price lunch (FRPL): low-poverty (0-40%), mid-poverty (40-75%), and high-poverty (over 75%). The second graph (b) presents the distribution of median incomes for a school's associated region.

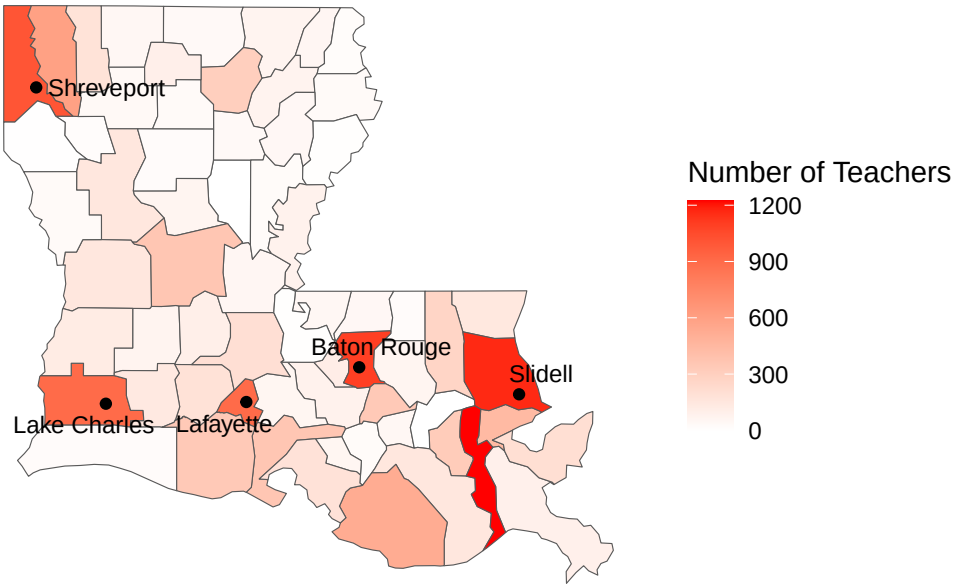



Figure A.5: Geographic distribution of Zearn teachers across parishes in Louisiana.

The color gradient represents the density of teachers, with darker hues indicating a higher concentration of educators using Zearn in each parish. The map also labels the top five cities where Zearn adoption is most prevalent.

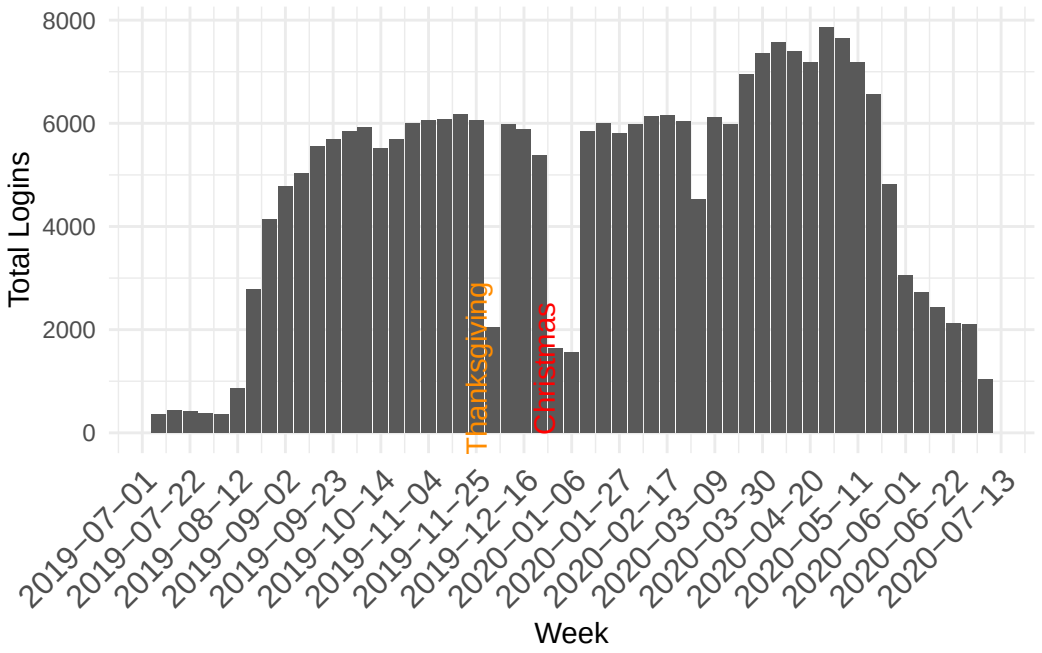



Figure A.6: Total number of student logins over the 2019-2020 school year.

The chart depicts the connection between academic schedules and platform engagement. Each bar represents a week, with peaks corresponding to active school weeks and troughs aligning with major holiday periods (e.g., Thanksgiving and Winter Break).

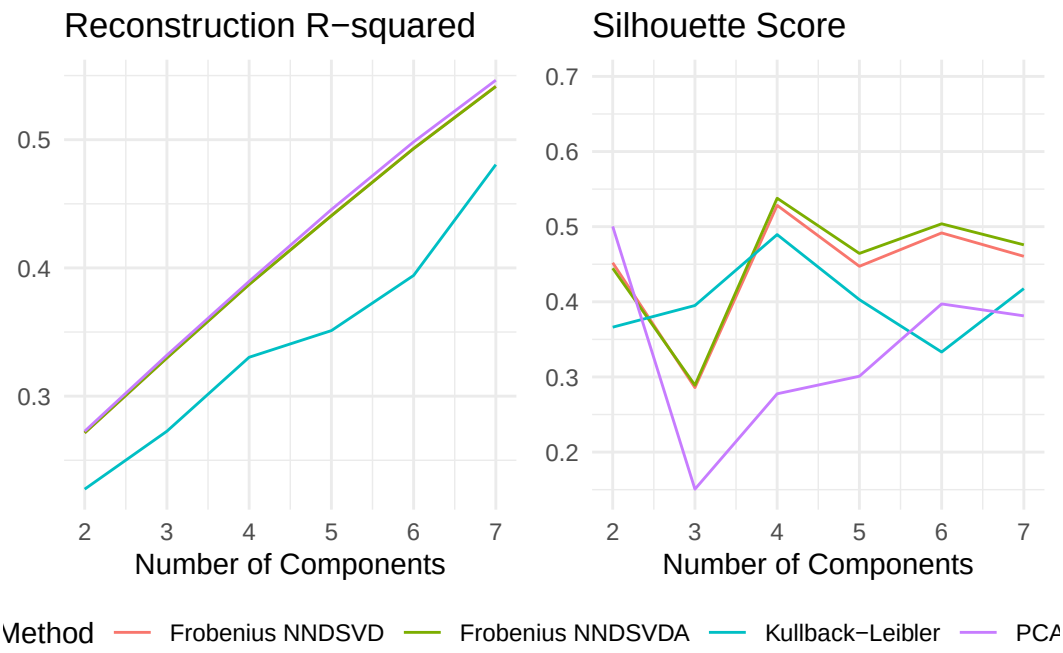



Figure A.7: Teacher data

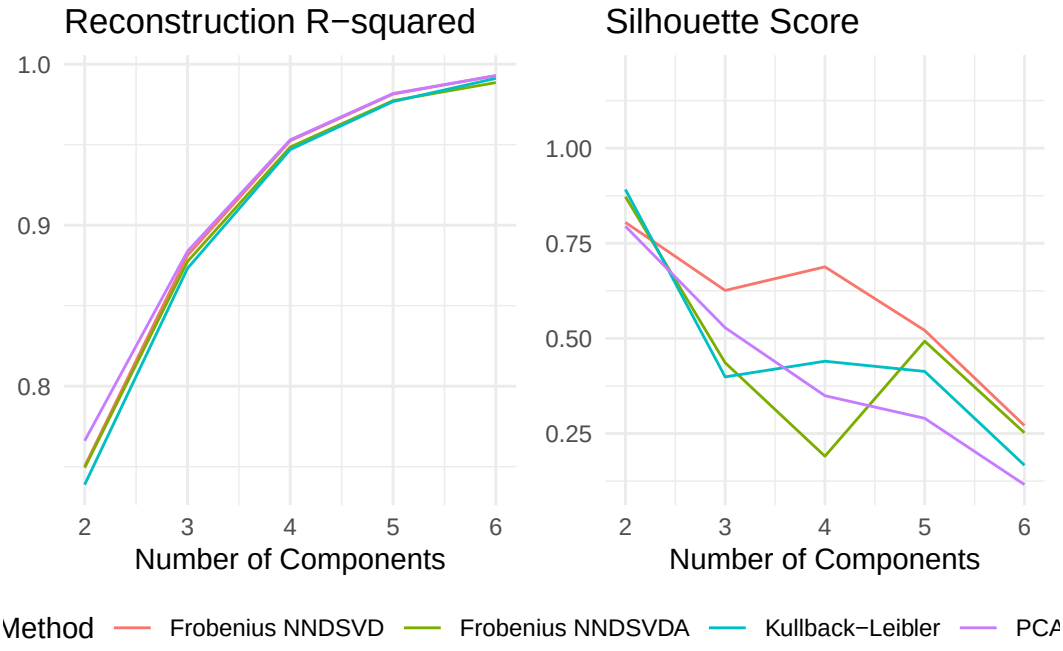



Figure A.8: Student data

Comparison of dimensionality reduction techniques for teacher and student data.

The figures compare the performance of Principal Component Analysis (PCA) against Nonnegative Matrix Factorization (NMF) in reducing the dimensionality of teacher and student data. The NMF variants include the Frobenius norm with two different initialization strategies: Nonnegative Double Singular Value Decomposition (Frobenius NNDSVD) and NNDSVD with the average of the input matrix X filled in place of zeros (Frobenius NNDSVDA). The third NMF variant uses the Kullback-Leibler divergence as the loss function. The left column assesses reconstruction quality using R-squared, where values closer to 1 indicate that the components can better recover the original data. The right column evaluates the interpretability of the low-dimensional representation using silhouette scores. Higher silhouette scores relate to better-defined clusters, values near 0 indicate overlapping clusters, and negative values generally suggest that a sample has been assigned to the wrong cluster.

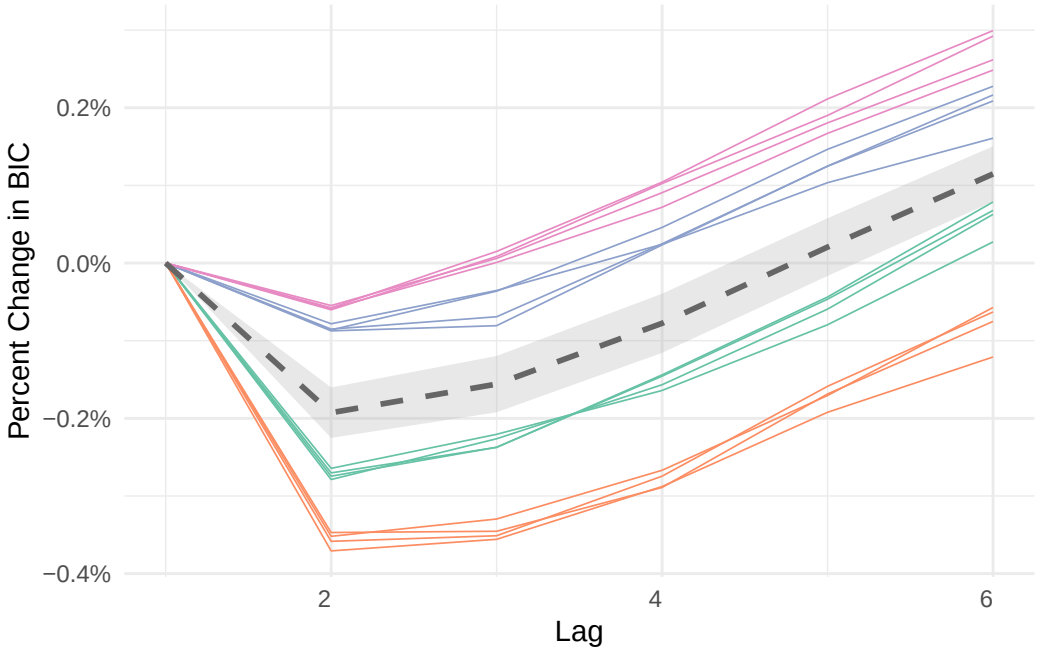



Figure A.9: BIC variations across lags for fixed-effects panel logistic regression models.

The plots show the percent change in model fit (BIC) for different lag periods compared to the one-week lag baseline. The thin lines represent the percent change for each combination of reward functions and methods, while the dashed gray lines represent their average. The shaded bands around the average lines indicate the standard error. The optimal lag period can be determined based on the minima in the BIC curves (lower BIC indicates better model fit when penalizing for complexity).

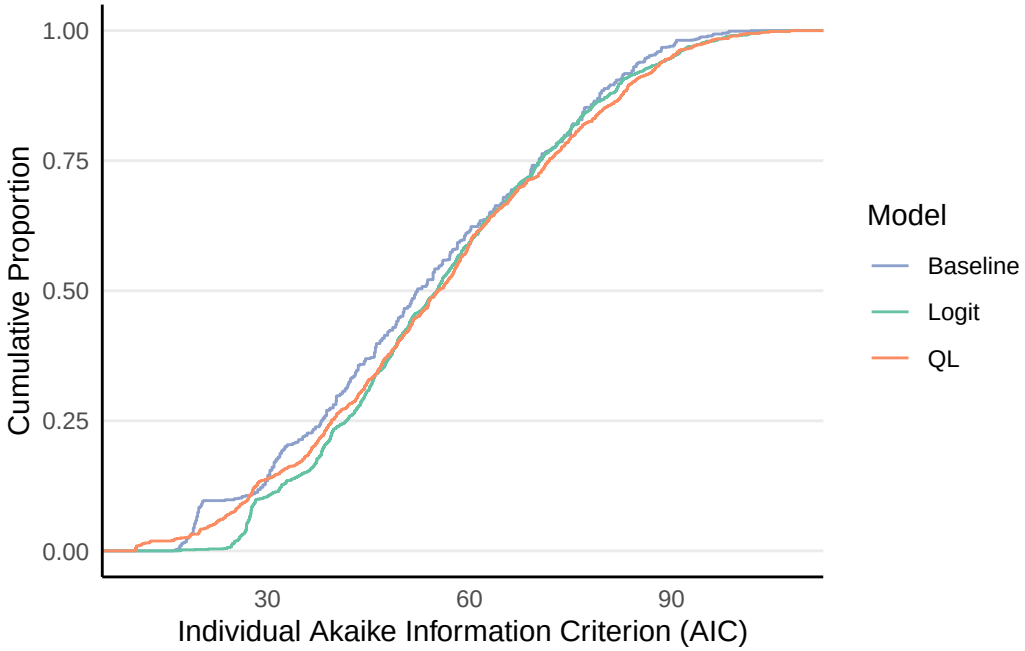
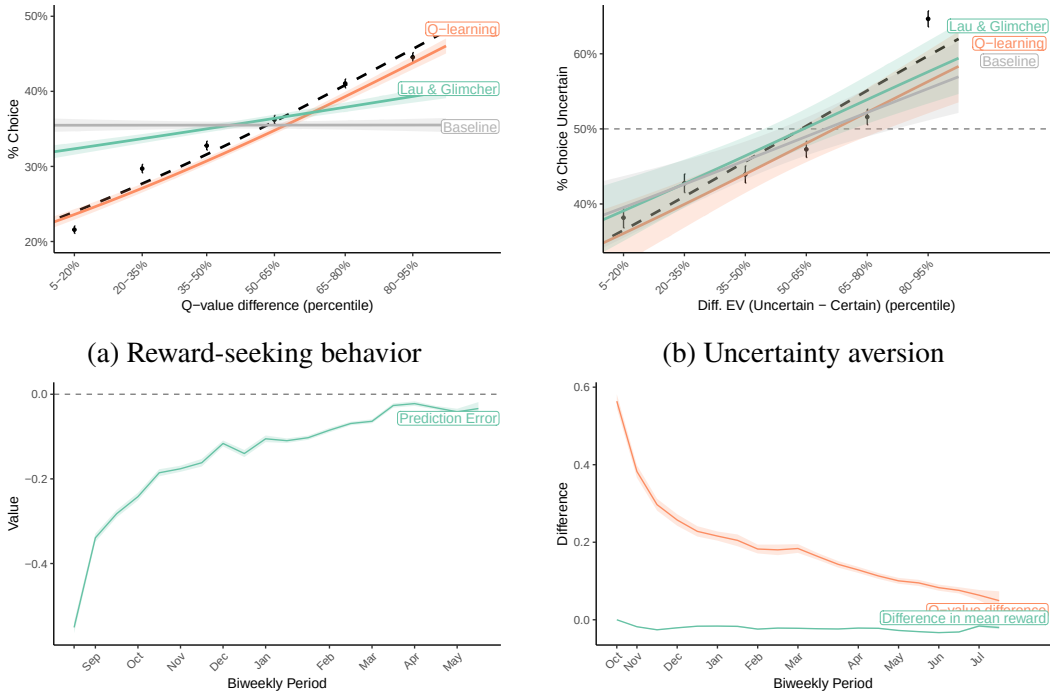



Figure A.10: Empirical Cumulative Distribution Function (ECDF) of teacher-specific Akaike Information Criteria (AIC) for different models.
Lower AIC values indicate better model fit. Models were estimated through non-hierarchical maximum likelihood estimation.



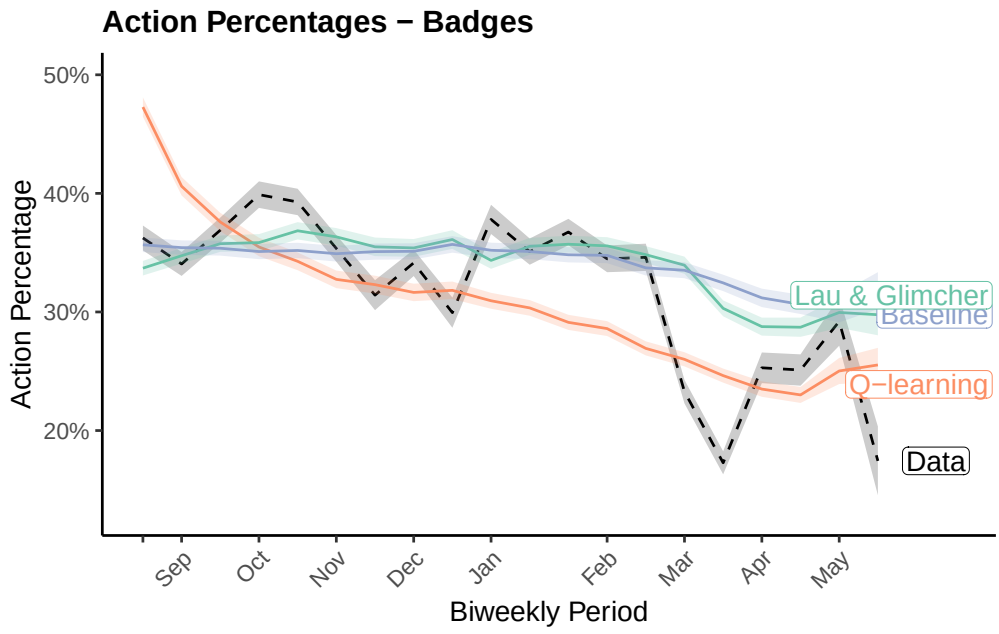

(a) Reward-seeking behavior

(b) Uncertainty aversion

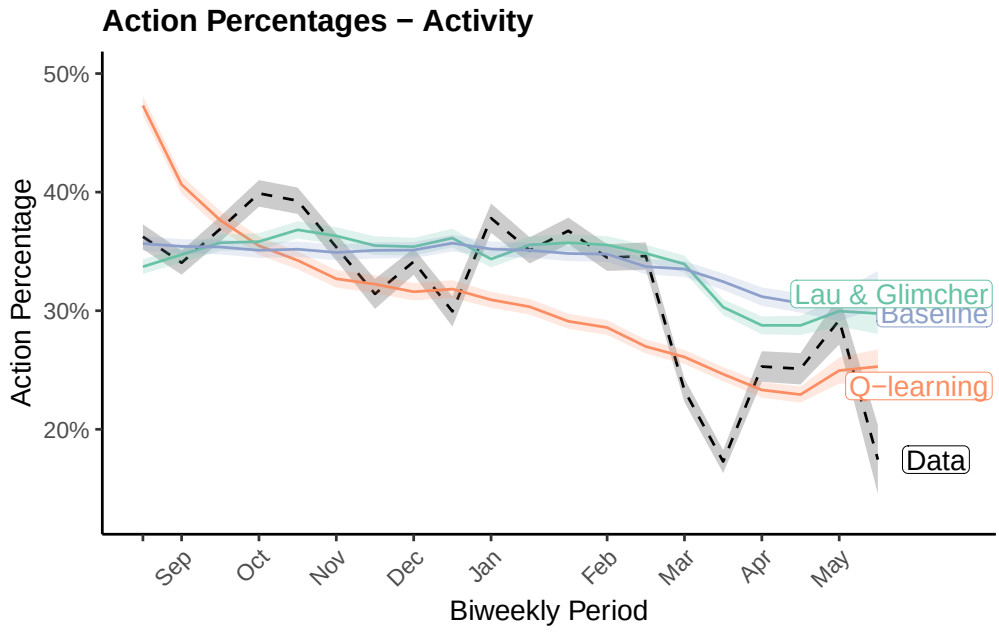
(c) Prediction Errors

(d) Q-value and reward differences

Figure A.11: Behavioral signatures of reinforcement learning in teacher decision-making for Activity as Rewards using non-hierarchical estimation.
The graphs compare three models (Q-learning, Logit, and Baseline) in their ability to capture various aspects of teachers' behavior. (a) Reward-seeking behavior: The x-axis represents the percentile of the difference in Q-values between action and inaction. The y-axis shows the proportion of times teachers chose to act. (b) Uncertainty aversion: The x-axis represents the percentile of the difference in expected value (EV) between uncertain and certain options, calculated from the cumulative means and standard deviations of rewards associated with each action. The y-axis shows the proportion of times teachers chose the uncertain option. (c) Prediction Errors: The plot shows the mean reward prediction errors across teachers over time. (d) Q-value and reward differences: The graph shows the difference in Q-values or mean rewards between action and inaction over time. In all plots, black points or dashed lines represent observed teacher behavior, while colored lines and shaded areas show model predictions with 95% confidence intervals.

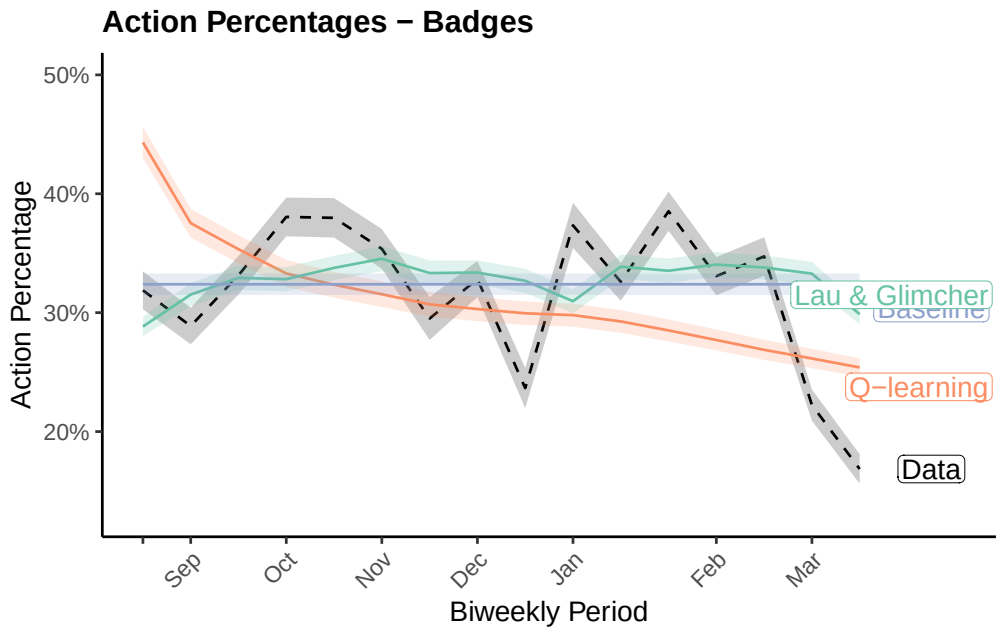



(a) Badges (unbalanced panel)

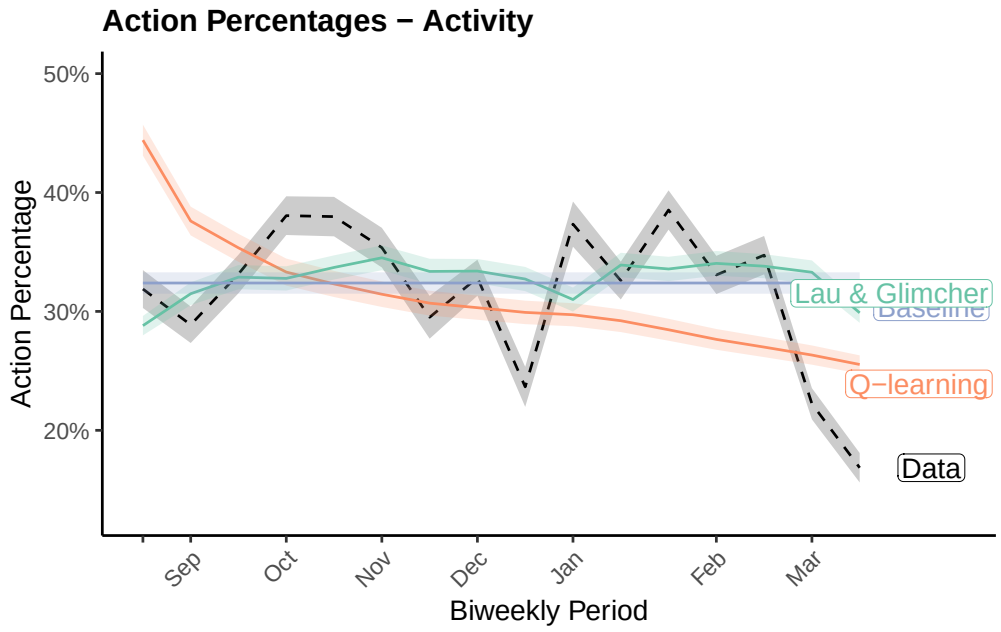



(b) Tower Alerts (unbalanced panel)

Figure A.12: Observed teacher behavior and model predictions over the academic year using the full, unbalanced dataset.
The graphs show the percentage of times teachers chose to engage in Pedagogical Content activities averaged across all teachers for each biweekly period. The x-axis represents biweekly periods throughout the school year. The black dashed line represents observed teacher behavior, while colored lines represent predictions from different models estimated through non-hierarchical maximum likelihood. Shaded areas represent the 95% confidence intervals for each model's predictions. Note that this unbalanced dataset includes teachers with missing data, resulting in varying numbers of teachers contributing to each time point, causing fluctuations in the average baseline predictions.




(a) Badges (balanced panel)




(b) Tower Alerts (balanced panel)

Figure A.13: Observed teacher behavior and model predictions over the academic year using a balanced subset of the data.
The graphs show the percentage of times teachers chose to engage in Pedagogical Content activities using only teachers with complete data for 16 biweekly periods (N = 632 teachers). The x-axis represents biweekly periods throughout the school year. The black dashed line represents observed teacher behavior, while colored lines represent predictions from different models estimated through non-hierarchical maximum likelihood. Shaded areas represent the 95% confidence intervals for each model's predictions. Note that the all teachers contribute to each time point equally, resulting in a baseline model with constant average predictions over time.

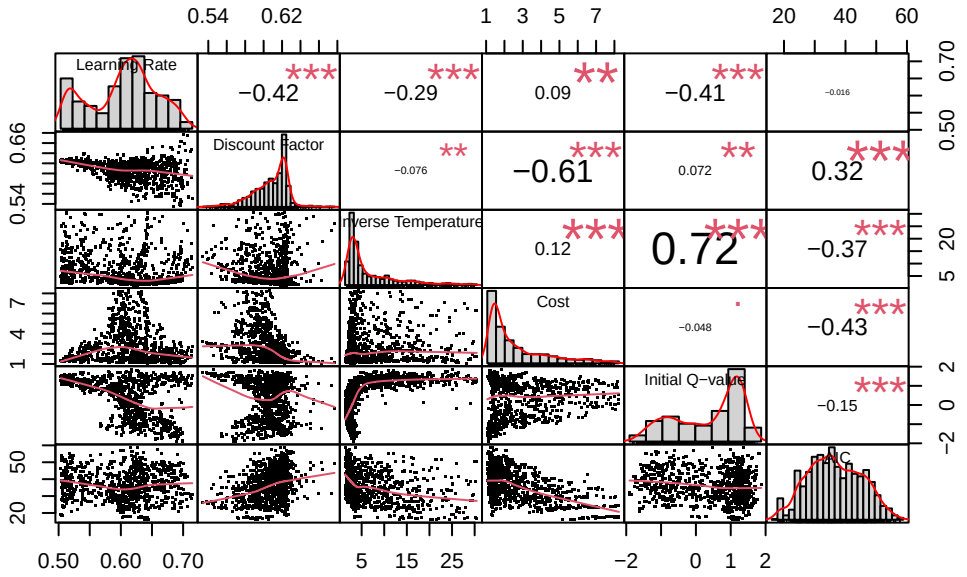



Figure A.14: Correlation matrix and distributions of reinforcement learning parameters and model fit.
The figure illustrates the pairwise Spearman correlations between key reinforcement learning parameters and model fit (AIC) derived from Q-learning model estimated through non-hierarchical maximum likelihood. The diagonal shows the distribution of each parameter, with histograms for discrete variables and density plots for continuous variables. The lower triangle displays scatterplots of pairwise relationships, with locally weighted scatterplot smoothing (LOWESS) lines in blue. The upper triangle presents correlation coefficients, with asterisks indicating statistical significance (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

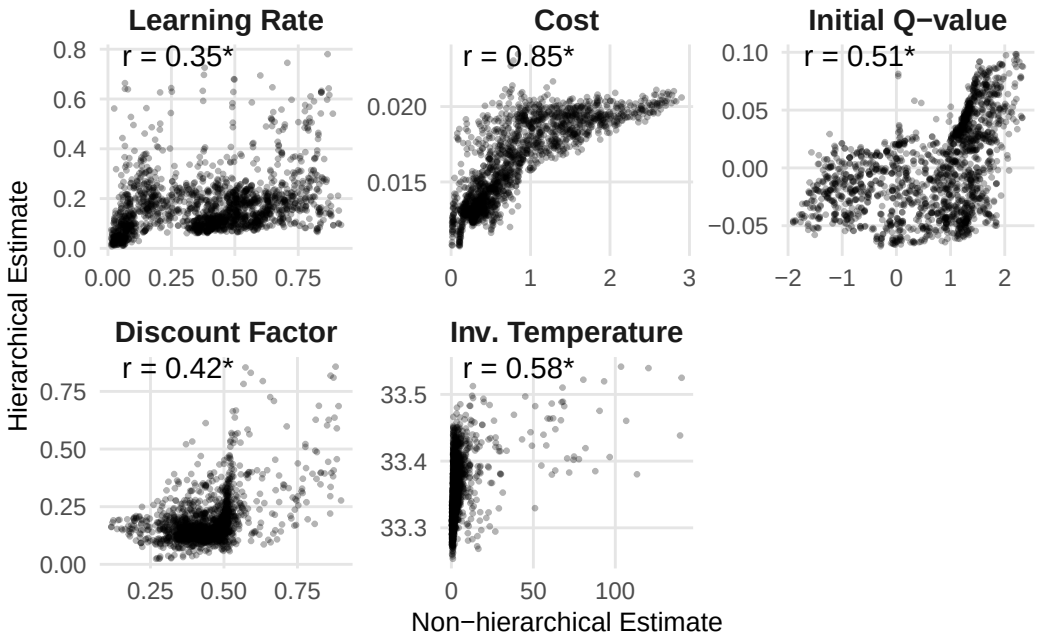



Figure A.15: Correlations between individual parameter estimates from non-hierarchical and hierarchical estimation approaches.
The figure illustrates the pairwise Spearman correlations between the two estimation methods. (*$p < 0.001$).

## References

Ding, C., He, X., & Simon, H. D. (2005). Proceedings of the 2005 siam international conference on data mining [518 citations (Crossref) [2024-03-22]], 606–610. https://doi.org/10.1137/1.9781611972757.70

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791. https://doi.org/10.1038/44565

Zearn. (2023). Curriculum study pd. https://webassets.zearn.org/Implementation/PDCourseCatalog.pdf

Zearn. (2024a). Adaptive fluency: Number gym. https://help.zearn.org/hc/en-us/articles/360052426453-Adaptive-Fluency-Number-Gym

Zearn. (2024b). Assessments. https://help.zearn.org/hc/en-us/articles/4403429312279-Assessments

Zearn. (2024c). Boosts. https://help.zearn.org/hc/en-us/articles/1500003390061-Boosts

Zearn. (2024d). Curriculum map. https://help.zearn.org/hc/en-us/articles/115016098288-Curriculum-Map

Zearn. (2024e). Guided practice. https://help.zearn.org/hc/en-us/articles/360052426573-Guided-Practice

Zearn. (2024f). Independent practice: Tower of power. https://help.zearn.org/hc/en-us/articles/360052426593-Independent-Practice-Tower-of-Power

Zearn. (2024g). Kindergarten activity tracker. https://help.zearn.org/hc/en-us/articles/1500000413782-Kindergarten-Activity-Tracker

Zearn. (2024h). Kindergarten digital activities. https://help.zearn.org/hc/en-us/articles/1500003414022-Kindergarten-Digital-Activities

Zearn. (2024i). Lesson materials. https://help.zearn.org/hc/en-us/articles/4403432269719-Lesson-Materials

Zearn. (2024j). Lesson-aligned fluency. https://help.zearn.org/hc/en-us/articles/236278807-Lesson-Aligned-Fluency

Zearn. (2024k). Math professional development. https://about.zearn.org/professional-development

Zearn. (2024l). Mission-level assessments. https://help.zearn.org/hc/en-us/articles/115007900988-Mission-level-assessments

Zearn. (2024m). Optional practice materials. https://help.zearn.org/hc/en-us/articles/4403429109143-Optional-Practice-Materials

Zearn. (2024n). Overview materials. https://help.zearn.org/hc/en-us/articles/4403432293271-Overview-Materials

Zearn. (2024o). Paper exit ticket. https://help.zearn.org/hc/en-us/articles/115007738707-Paper-Exit-Ticket

Zearn. (2024p). Teaching with zearn math in louisiana. https://help.zearn.org/hc/en-us/articles/1500009157742-Teaching-with-Zearn-Math-in-Louisiana

Zearn. (2024q). Use zearn as homework. https://help.zearn.org/hc/en-us/articles/115007901328-Use-Zearn-as-homework

Zearn. (2024r). Zearn math for kindergarten. https://help.zearn.org/hc/en-us/articles/115008330968-Zearn-Math-for-Kindergarten

Zearn. (2024s). Zearn recommended schedule. https://help.zearn.org/hc/en-us/articles/236278407-Zearn-Recommended-Schedule

Zearn. (2024t). Zearn student notes. https://help.zearn.org/hc/en-us/articles/5255954252311-Zearn-Student-Notes

*Appendix B*

# APPENDIX FOR CHAPTER 3

## B.1 Supplementary Methods

### Differences in the Original vs. Revised Analysis

As noted in the main text, we have corrected a few inaccuracies in our initial analysis, improving the robustness of our methodology. To be fully transparent, we have outlined below all the changes made from the original to the revised analysis. Please refer to Tables S3 to S5 in this supplementary document for the original results that inspired our treatments.

1. Temporal Alignment Adjustments: The original method used the standard week and year delineation based on the Gregorian calendar. The new analysis uses the ISO week date system to ensure correct week numbering, particularly around the transition from one year to the next, which corrected previous week misclassifications.

2. Duplicate Record Management: The dataset contains a small number of duplicate classroom-week pairs generated in classrooms with more than one teacher linked to it. The original method removed the first duplicate occurrence, consequently discarding data from teachers with larger ID numbers. The revised analysis orders and filters duplicate records by the number of classes each teacher manages, retaining data from teachers involved in fewer classes, thereby minimizing the inclusion of supervisory rather than direct instructional roles.

3. Independent Component Analysis: The original ICA used the `fastICA` package, presenting some inconsistencies: non-deterministic component sequencing, arbitrary sign inversions, and varied loading coefficients, even with a set seed. The updated approach adopts the `ica` package, ensuring orderly component arrangement and consistent outputs. Sign orientation is now standardized, maintaining the largest loading variable as positive. Further, for the models restricted to schools that use Zearn as a main component of their curriculum, the ICA was estimated separately (see Table S4).

4. Model Selection Adjustment: Initially, we estimated a random effects model. Following a Hausman test indicating inconsistency (chi-square = 81.31, df = 4, p < .001), we transitioned to a fixed effects model. The revised approach also incorporates robust standard errors, adjusting for heteroskedasticity and autocorrelation (Arellano, 1987; Long & Ervin, 2000).

5. Removed User.Session: The original ICA included "User Session." This variable measures the frequency of a teacher's logins to the Zearn platform. As such, it does not offer substantive insight into the pedagogical nature of the teachers' interactions with the platform. We aimed to understand the educational impact of specific usage patterns rather than their frequency. Hence, "User Session" was deemed a nuisance variable as it risked overshadowing more pertinent patterns related to instructional engagement and effective pedagogical strategies. In particular, this variable mainly loads onto the "empathy" component. Its exclusion in our revised analysis has confirmed that the identified patterns genuinely reflect empathy in teaching approaches (see Tables S3 and S4).

**Independent Component Analysis (ICA)**

We implement the FastICA algorithm (Hyvärinen & Oja, 2000) to estimate independent components from our dataset. In this model, matrix $X = \{x_{ij}\}_{I\times J}$ , consisting of $I$ samples across $J$ random variables, is expressed as a linear mixture of independent components $C$, represented by:

$$X = C'M + E.$$

Here, $C$ holds the independent components, $M$ is a mixing matrix, and $E$ denotes the noise. The aim is to minimize mutual information between components in $C$, which is achieved by maximizing their marginal negentropy, thereby rendering the columns of $C$ statistically independent.

The FastICA process begins by transforming $X$ into a whitened matrix $Y$, ensuring uncorrelated variables with unit variance. This transformation is achieved through eigenvalue decomposition:

$$Y = X \cdot \left[ \frac{\mathbf{v_1}}{\sqrt{\lambda_1}} \quad \frac{\mathbf{v_2}}{\sqrt{\lambda_2}} \quad \frac{\mathbf{v_3}}{\sqrt{\lambda_3}} \right],$$

where $(\lambda_1, \lambda_2, \lambda_3)$ and $(\mathbf{v_1}, \mathbf{v_2}, \mathbf{v_3})$ are, respectively, the eigenvalues and eigenvectors of $\frac{X'X}{I}$.

Afterward, the algorithm approximates the negentropy with

$$\hat{\theta}(c_n) = (\mathbb{E}[\ln(\cosh(c_n))] - \mathbb{E}[\ln(\cosh(z))])^2$$

where $c_n, n \in 1, 2, 3$, is one of the components, and $z$ is a Gaussian variable with zero mean and unit variance. FastICA iteratively maximizes this value across all components, producing an orthogonal rotation matrix $R_{3\times3}$ such that $C = YR$.

For more details on ICA and the FastICA algorithm, see Hyvärinen and Oja (2000) and Helwig and Hong (2013).

**Focus Group Discussions**

In 2021, we conducted regular meetings (once a month, on average) with Zearn employees to discuss the interpretation of our data analyses. Additionally, on April 6th and 13th, 2021, we held "office hours" with Zearn teachers and administrators. For these specific meeting we prepared the following questions, although due to time limitations, we were unable to ask them systematically:

1. With what regularity do you find logging into Zearn most helpful?
2. What tasks do you typically do on Zearn?
3. Please order the tasks you mentioned in how important they are in helping your teaching (from most important to least important).
4. Please explain why [top option from question 3] is the most important.
5. Have you done activities and exercises designed for the students on Zearn?
6. If so, with what regularity do you do activities and exercises designed for the students on Zearn?

Below are some of the most relevant quotes from these meetings:

"We could layer in some of the insights here, [for example], prompting teachers to go and do towers and get towers and go through mediation paths. I think that could be something we layer into one of those A/B tests." - B.M., Zearn administrator

"If students are not successful in Tower of Power, they get stuck, and it is the loop of self defeat - I know what I need to do, but I can't do it. Teachers don't check the tower alerts report enough so they can defeat that cycle." - Math coach from CA

"For everything I do, I try to look through the kid's eyes. I look for the lightbulb moment, and the thing that finally gets kids to understand it." -5th grade teacher from IL

"One thing is when kids really understand something and have that "aha" moment - and you know you have gotten through. And that could be academically or socially." - Middle school teacher from CA

"It took my students 50-60 minutes to get through a lesson. I have some kids who get fatigued by the length of the digital instruction and program. The boosts may not help them. Need to orient them to the Zearn lesson." - Zearn Teacher

## B.2 Supplementary Tables

Table B.1: Independent Component Analysis (ICA) Results.

This table displays the weights of each teacher activity in the three independent components (ICs). Notably, these metrics pertain to teacher activity on the platform, not student actions.

| **Activity** | **IC 1** | **IC 2** | **IC 3** |
|---|---|---|---|
| Tower Struggled | 0.89 | -0.02 | 0.02 |
| Tower Stage Failed | 0.89 | -0.01 | 0.02 |
| Fluency Completed | 0.84 | 0.02 | 0.00 |
| Guided Practice Completed | 0.71 | 0.04 | -0.02 |
| Number Gym Activity Completed | 0.66 | 0.02 | 0.02 |
| Tower Completed | 0.65 | 0.09 | 0.02 |
| Grade Level Overview RD | 0.12 | 0.16 | -0.35 |
| Student Notes and Exit Tickets RD | 0.10 | 0.06 | -0.34 |
| Kindergarten Activity Completed | 0.06 | -0.01 | -0.01 |
| Mission Overview RD | 0.03 | 0.37 | 0.17 |
| Kindergarten Schedule RD | 0.03 | -0.00 | 0.02 |
| Teaching and Learning Approach RD | 0.02 | -0.17 | 0.32 |
| Optional Problem Sets RD | 0.01 | 0.65 | 0.02 |
| Optional Homework RD | 0.01 | 0.67 | 0.10 |
| Whole Group Word Problems RD | 0.01 | 0.67 | 0.10 |
| Elementary Schedule RD | 0.00 | 0.02 | 0.28 |
| Small Group Lesson RD | 0.00 | 0.72 | 0.06 |
| Curriculum Map RD | 0.00 | 0.00 | 0.31 |
| Kindergarten Mission RD | 0.00 | -0.03 | 0.23 |
| Assessments RD | -0.01 | 0.55 | 0.08 |
| PD Course Guide RD | -0.01 | -0.03 | 0.66 |
| Whole Group Fluency RD | -0.01 | 0.64 | 0.15 |
| Assessments Answer Key RD | -0.01 | 0.45 | 0.14 |
| PD Course Notes RD | -0.01 | 0.08 | 0.66 |

Table B.2: Marginal effects of ICA components on badges.

Marginal effects are calculated by multiplying the ICA coefficients by the mean of each component and dividing by the standard deviation of the corresponding variable.

| **Activity** | **Effect of 1 SD** | **Effect of 1 Unit** |
|---|---|---|
| Tower Struggled | 4.94 | 7.14 |
| Tower Stage Failed | 4.89 | 2.03 |
| Fluency Completed | 4.74 | 3.47 |
| Guided Practice Completed | 4.10 | 2.69 |
| Number Gym Activity Completed | 3.70 | 5.63 |
| Tower Completed | 3.66 | 2.73 |
| Student Notes and Exit Tickets RD | 0.59 | 0.88 |
| Kindergarten Activity Completed | 0.17 | 0.73 |
| Optional Homework RD | 0.04 | 0.06 |
| Optional Problem Sets RD | 0.02 | 0.02 |
| Small Group Lesson RD | -0.77 | -0.81 |
| Assessments RD | -0.89 | -1.76 |
| Whole Group Word Problems RD | -1.05 | -1.70 |
| Mission Overview RD | -1.18 | -2.03 |
| Assessments Answer Key RD | -1.24 | -3.71 |
| Whole Group Fluency RD | -1.45 | -2.84 |
| Kindergarten Mission RD | -1.54 | -5.43 |
| Kindergarten Schedule RD | -1.71 | -84.55 |
| Grade Level Overview RD | -1.79 | -13.01 |
| Elementary Schedule RD | -1.87 | -34.91 |
| Teaching and Learning Approach RD | -2.09 | -41.00 |
| Curriculum Map RD | -2.09 | -46.13 |
| PD Course Guide RD | -4.47 | -67.43 |
| PD Course Notes RD | -4.51 | -31.16 |

## B.3 Supplementary Figures

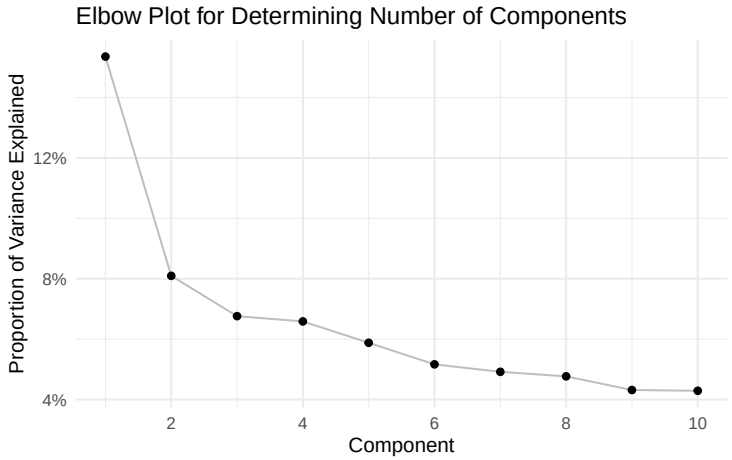



Figure B.1: Elbow (Scree) plot for determining optimal number of components.

The plot displays the proportion of variance explained by each independent component (IC). The optimal number of components is indicated by the "elbow" of the plot, where the variance explained by each additional component is minimal.

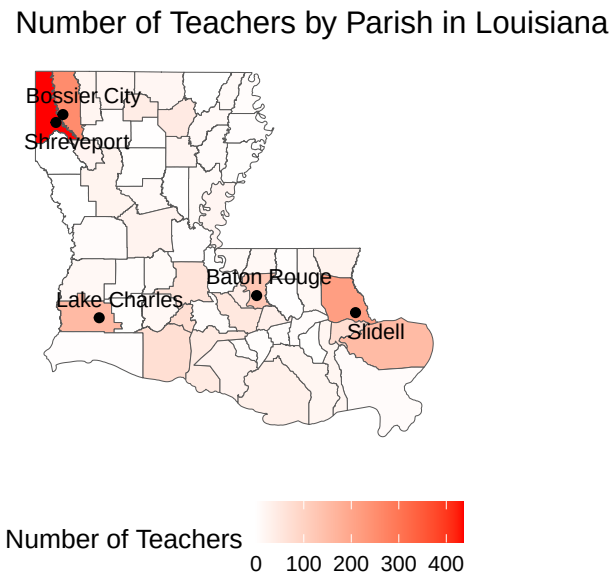



Figure B.2: Geographical distribution of teachers across various parishes in Louisiana, and the top 5 cities with the highest number of teachers.

## Original Analysis

Table B.3: Independent Component Analysis (ICA) results without including user sessions.

| Activity | IC 1 | IC 2 | IC 3 |
|---|---|---|---|
| Tower Struggled | 0.89 | -0.03 | 0.02 |
| Tower Stage Failed | 0.88 | -0.02 | 0.03 |
| Fluency Completed | 0.83 | 0.00 | 0.00 |
| Guided Practice Completed | 0.71 | 0.04 | -0.02 |
| Number Gym Activity Completed | 0.66 | -0.01 | 0.01 |
| Tower Completed | 0.65 | 0.09 | 0.00 |
| User Session | 0.54 | 0.38 | -0.01 |
| Grade Level Overview RD | 0.11 | 0.16 | -0.36 |
| Student Notes and Exit Tickets RD | 0.11 | 0.14 | 0.04 |
| Kindergarten Activity Completed | 0.06 | -0.01 | -0.01 |
| Mission Overview RD | 0.03 | 0.37 | 0.18 |
| Kindergarten Schedule RD | 0.03 | -0.05 | 0.28 |
| Teaching and Learning Approach RD | 0.02 | -0.17 | 0.32 |
| Optional Problem Sets RD | 0.02 | -0.65 | -0.05 |
| Optional Homework RD | 0.01 | 0.60 | -0.05 |
| Small Group Lesson RD | 0.01 | 0.72 | 0.06 |
| Elementary Schedule RD | 0.00 | 0.02 | 0.28 |
| Curriculum Map RD | 0.00 | 0.00 | 0.31 |
| Whole Group Word Problems RD | 0.00 | 0.66 | 0.11 |
| Kindergarten Mission RD | 0.00 | -0.03 | 0.23 |
| PD Course Guide RD | 0.00 | -0.02 | 0.66 |
| Assessments RD | -0.01 | 0.54 | 0.09 |
| PD Course Notes RD | -0.01 | 0.07 | 0.65 |
| Whole Group Fluency RD | -0.02 | 0.63 | 0.17 |
| Assessments Answer Key RD | -0.02 | 0.44 | 0.15 |

Variance Accounted For: 15.75%, 12.49%, 6.06 %

## References

Arellano, M. (1987). Practitioners' corner: Computing robust standard errors for within-groups estimators * [876 citations (Crossref) [2024-06-04]]. *Oxford Bulletin of Economics and Statistics*, *49*(4), 431–434. https://doi.org/10.1111/j.1468-0084.1987.mp49004006.x

Helwig, N. E., & Hong, S. (2013). A critique of tensor probabilistic independent component analysis: Implications and recommendations for multi-subject fmri data analysis [24 citations (Crossref) [2023-12-27]]. *Journal of Neuroscience Methods*, *213*(2), 263–273. https://doi.org/10.1016/j.jneumeth.2012.12.009

Hyvärinen, A., & Oja, E. (2000). Independent component analysis: Algorithms and applications [5771 citations (Crossref) [2023-12-27]]. *Neural Networks*, *13*(4-5), 411–430. https://doi.org/10.1016/S0893-6080(00)00026-5

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model [299 citations (Crossref) [2024-06-04]]. *The American Statistician*, *54*(3), 217–224. https://doi.org/10.1080/00031305.2000.10474549

*Appendix C*

# APPENDIX FOR CHAPTER 4

## C.1 Statistical Definitions

### The random effects model

The random-effects model accounts for between-study heterogeneity, which causes true effect sizes of studies to differ. Specifically, the model assume that $\hat{\theta}_k = \theta_k + \epsilon_k$, where $\hat{\theta}_k$ is the observed effect size, $\theta_k$ is the true effect size of study $k$, and $\epsilon_k$ is the sampling error. Furthermore, it assumes that the true effect size $\theta_k$ of study $k$ is only part of an over-arching distribution of true effect sizes with mean $\mu$: $\theta_k = \mu + \zeta_k$. Overall the random-effects model can be expressed as $\theta_k = \mu + \zeta_k + \epsilon_k$, indicating that the observed effect size deviates from the pooled effect $\mu$ because of two error terms, $\zeta_k$ and $\epsilon_k$.

The pooled effect size is a weighted average of all studies. The weight $w_k$ for each study $k$ is calculated as the *inverse-variance*: $w_k^* = \frac{1}{s_k^2 + \hat{\tau}^2}$, where $s_k^2$ is the estimated within-study variance of the observed effect size, capturing the variability in study outcomes due to sampling error, and $\hat{\tau}^2$ is the estimated between-study variance, which accounts for the true effect sizes' heterogeneity across different studies.

To correct for our small samples, we adjust the model via the Knapp–Hartung modification, also known as the Sidik–Jonkman modification, (Hartung & Knapp, 2001a, 2001b; Sidik & Jonkman, 2002), which, unlike the more common restricted maximum likelihood (REML) estimation, does not assume that the error distribution is normal. This technique adjusts the standard errors of the regression coefficients (including the intercept-only model, which calculates the meta-analytic effect size) by multiplying their variances by $q_{\mathrm{KH}} = \frac{\widehat{\boldsymbol{\theta}}\mathbf{P}\widehat{\boldsymbol{\theta}}}{K-p}$, with $\mathbf{P} = \mathbf{W}^* - \mathbf{W}^*\mathbf{X}\left(\mathbf{X}'\mathbf{W}^*\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}^*$, where $\mathbf{W}^* = \operatorname{diag}\left(w_1^*, w_2^*, \ldots, w_K^*\right)$, and $\mathbf{X} = \left(\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_K'\right)'$ is the matrix of a vector from each study $k \in K$ of $p$ moderators (including the intercept). Intuitively, this method incorporates the uncertainty of estimating $\tau^2$, a factor that increases with smaller number of studies.

### Correlation as an effect size

For our analysis with names, we deploy the random-effects model to estimate the true correlation between PC1 and callback, $\rho$, with $r_k(\text{callback}, \text{PC1})$. The value

$r_k$ is transformed into Fisher's $z_k$: $z_k = 0.5 \log_e \left(\frac{1+r_k}{1-r_k}\right)$, to ensure that the sampling distribution is approximately normal.

**The between-study variance, $\tau^2$**

To estimate the random-effects model, the error $\zeta_k$ must be considered. To do this, the variance of the distribution of true effect sizes, $\tau^2$, has to be estimated. There are several methods to estimate $\tau^2$, we deploy Maximum Likelihood (Viechtbauer, 2005).

The confidence intervals around $\tau^2$ can be estimated using various methods, which depend on the type of $\tau^2$ estimator used. We deploy the *Q-Profile* method Veroniki et al., 2016 which is based on the $Q_{gen}$ statistic: $Q_{gen} = K \sum_{k=1}^{k} w_k^*(\hat{\theta}_k - \hat{\mu})^2$. The Q-Profile method uses an iterative process: $Q_{gen}(\tilde{\tau}^2)$ is calculated repeatedly while increasing the value of $\tau^2$ until the expected value of the lower and upper bound of the confidence interval based on the $\chi^2$ distribution is reached.

Unlike the standard version of Q, which uses the pooled effect based on the fixed-effect model, Qgen is based on the random-effects model and uses the overall effect according to this model, denoted by $\hat{\mu}$, to calculate the deviates. Additionally, Qgen uses weights based on the random-effects model, denoted by $w_k^*$, in its calculation. The equation for Qgen is given by $Q_{gen} = \sum_{k=1}^{K} w_k^* \left(\hat{\theta}_k - \hat{\mu}\right)^2$, where $w_k^*$ is the random-effects weight $w_k^* = \frac{1}{s_k^2+\tau^2}$.

**The hetereogeneity measure, $I^2$**

We calculate the $I^2$ statistic Thompson and Higgins, 2002 to provide an estimate of the magnitude of the between-study heterogeneity. $I^2$ represents the percentage of the total variability in the effect sizes not due to sampling error, formally expressed as $I^2 = \frac{Q-(K-1)}{Q}$, where $K$ is the total number of studies. Cochran's $Q$ Cochran, 1954 is defined as $Q = \sum_{k=1}^{K} w_k(\hat{\theta}_k - \hat{\theta})^2$. It uses the deviation of each study's observed effect $\hat{\theta}_k$ from the summary effect $\hat{\theta}$, weighted by the inverse of the study's variance, $w_k$. The test statistic for Cochran's $Q$ is distributed as chi-squared with $K-1$ degrees of freedom under the null hypothesis of homogeneity. The value of $I^2$ cannot be lower than 0%. If $Q$ is smaller than $K-1$, 0 is used instead of a negative value.

**Prediction intervals**

Prediction intervals provide a valuable tool for estimating the likely range of effects that future studies may have based on the current evidence. The formula for 95%

prediction intervals according to IntHout et al., 2016, is calculated as follows:
$\hat{\mu} \pm t_{K-1,0.975}\sqrt{SE^2_{\hat{\mu}} + \hat{\tau}^2}$
$\hat{\mu} \pm t_{K-1,0.975} SD_{PI}$, where $K$ is the number of studies.

**Mixed-effects models**

The meta-regressions were specified as mixed-effects models: $\hat{\theta}_k = \theta + \beta x_k + \epsilon_k + \zeta_k$. The first error, $\epsilon_k$, represents the sampling error through which a study's effect size deviates from its true effect. The second error, $\zeta_k$, indicates that even the true effect size of a study is only sampled from an overarching distribution of effect sizes.

**Intraclass correlation (ICC)**

We calculated the ICC through a two-way random-effects model (as provided by package `psych`) to assess the reliability of the average of $k$ ratings for each signal $i$. We describe each rating as $y_{ij} = \mu + r_i + c_j + e_{ij}$, where $\mu$ is the average rating, $r_i \sim N(0, \sigma_r^2)$ and $c_j \sim N(0, \sigma_c^2)$ are random effects for the signals and raters, respectively, and $e_{ij}$ is the error term. Then, we compute ICC $= \frac{\sigma_r^2}{\sigma_r^2 + \left(\sigma_c^2 + \sigma_\epsilon^2\right)/k}$ Liljequist et al., 2019.

## C.2 Supplementary Methods

**Heterogeneity analysis**

On our main meta-model, we computed several influence diagnostics (Externally Standardized Residuals, DFFITS Value, Cook's Distance, Covariance Ratio, Leave-One-Out $\tau^2$, Hat Value, Study Weight), which did not nominate any study as an outlierFig. C.3. Additionally, we implemented a Graphical display of heterogeneity (GOSH) plot analysis (C.5). For this analysis, we fit all possible subsets $2^{k-1}$ of our included studies. And plot the pooled effect size against the between-study heterogeneity. Three (k-means, DBSCAN, gmm) clustering algorithms are used to determine patterns. Two (DBSCAN, GMM) algorithms detected the same potential outliers: Jacquemet and Yannelis, 2012; Kline et al., 2022b; Neumark et al., 2019; Nunley et al., 2017. Excluding those yields $\theta = .22$, $p$-value=.19.

Furthermore, we visualized the contribution of each study to the overall heterogeneity against its influence on the pooled effect size (also known as Baujat plot, C.2). Kline et al., 2022b showed the highest contribution to heterogeneity, however, its influence on the pooled result was small. Neumark et al., 2019 showed a moderate contribution to the overall heterogeneity but a substantial influence on the pooled result. A leave-one-out robustness analysis also indicated that excluding Neumark

et al., 2019 resulted in the largest decrease in the $I^2$ statistic, reducing it from 81% to 63.8%. $\theta = .34$, $p$-value=.026 (C.4).

**Influence diagnostics**

In the following, we define the measures plotted in Fig. C.3. Fig. C.3 first panel, displays the *externally standardized residual* of each study, defined as follows:

$$t_k = \frac{\hat{\theta}_k - \hat{\mu}_{\setminus k}}{\sqrt{\text{Var}(\hat{\mu}_{\setminus k}) + \hat{\tau}^2_{\setminus k} + s^2_k}}. \tag{C.1}$$

These residuals are the deviation of each observed effect size $\hat{\theta}_k$ from the pooled effect size. The "external" pooled effect $\hat{\mu}_{\setminus k}$ is obtained by calculating the overall effect without study $k$. The resulting residual is then standardized by (1) the variance of the external effect $\hat{\mu}_{\setminus k}$), (3) the $\tau^2$ estimate of the external pooled effect, and (3) the variance of $k$.

Fig. C.3, second panel, displays the $\text{DFFITS}_k$. The DFFITS value indicates how much the pooled effect changes when a study $k$ is removed, expressed in standard deviations. Higher values indicate that a study may be influential because its impact on the average effect is larger.

$$DFFITS_k = \frac{\hat{\mu} - \hat{\mu}_{\setminus k}}{\sqrt{\frac{w^*_k}{\sum_{k=1}^{K} w^*_k}(\hat{s}^2_k + \hat{\tau}^2_{\setminus k})}} \tag{C.2}$$

where $w^*_k$ is the (random-effects) weight of study $k$.

Fig. C.3, third panel displays the Cook's distance value $D_k$ of a study. $D_k$ only takes positive values and is calculated as follows:

$$D_k = \frac{(\hat{\mu} - \hat{\mu}_{\setminus k})^2}{\sqrt{\hat{s}^2_k + \hat{\tau}^2}}. \tag{C.3}$$

Fig. C.3, fourth panel displays $\text{CovRatio}_k$. A value below 1 indicates that removing study $k$ results in a more precise estimate of the pooled effect size $\hat{\mu}$.

$$\text{CovRatio}_k = \frac{\text{Var}(\hat{\mu}_{\setminus k})}{\text{Var}(\hat{\mu})} \tag{C.4}$$

Fig. C.3, fifth and sixth panels display Leave-One-Out $\tau^2$ and $Q$ values. The values display the estimated heterogeneity as measured by $\tau^2$ and Cochran's $Q$ if study $k$ is

removed. Lower values of $Q$, but particularly of $\tau^2$, are desirable since this indicates lower heterogeneity.

Fig. C.3, seventh and eigth panels display the study weight and hat value of each study. The hat value is another metric that is equivalent to the study weight.

**Subgroup analysis: Social perceptions across job types**

This section looks further into studies that have published their full datasets. In the following analyses, each row corresponds to one CV sent out in the experiment. Columns for each study vary, but in all of them, we have access to the name on the CV and a variety of CV-specific characteristics (e.g., education level, previous experience).

We assume that the explanatory power of social perceptions is dependent on job-specific context. A base model (logit, no classes) would explain callbacks as such:

$$\Pr(\text{Callback} = 1) = \frac{\exp(\alpha + \beta \text{PC1})}{1 + \exp(\alpha + \beta \text{PC1})}$$

where $\alpha$ is the intercept and $\beta$ is the coefficient from this estimation. Thus, we compare this model with one where we run interactions with job types. We infer interactions between industries, occupations, and educational levels from six correspondence studies by employing finite mixture models (FMMs) for clustering occupations. FMMs allow multiple latent classes to reveal a more detailed relationship between stereotypes and callbacks across different job types.

FMMs are employed because they allow for the modeling of unobserved heterogeneity by identifying subgroups or clusters within the data. In this context, FMMs are used to identify different classes of occupations that have a similar relationship between callbacks and social perception features.

Mathematically, the FMM can be expressed as:

$$f(y_i) = \sum_{j=1}^{k} \pi_j f_j(y_i)$$

where is the $\pi_j$ is the mixing proportion for the $j$-th class, $f(y_i)$ the conditional probability density function for the observed response $y_i$ (callbacks) in the $i$-th class model, and $k = 2$ is the number of latent classes or clusters. In other words, we

estimate a different $\beta$ for each latent class. The softmax function determines latent class probability:

$$\pi_i = \frac{\exp(\gamma_i)}{\sum_{j=1}^{g} \exp(\gamma_j)} \tag{C.5}$$

where $\gamma_i$ is a function of job characteristics. The effect of PC1 on job characteristics was operationalized using text data from job advertisement titles and descriptions and other relevant variables such as industry, occupation, and education level. Industry and occupation variables provided insights into specific sectors and job roles, respectively, while the education level variable captured employers' educational requirements or preferences for certain positions. These additional variables were provided in the published datasets and offered contextual information about the jobs being studied. The dataset was preprocessed and cleaned to eliminate rare industries and occupations, retaining only those that appeared in at least 1% of the observations.

The job description text was parsed, and common words were extracted using the n-gram technique with n=1 (unigrams), which allowed for identifying meaningful patterns in the text data (Jurafsky & Martin, 2023). Unigrams are single words from the text, which help capture the frequency of individual words and their potential importance in characterizing jobs.

For each study, we fit a base model with no latent classes and compare it to separate FMMs with two classes. We then evaluate the fit of these models by comparing their Bayesian information criterion (BIC) values to determine the best model for each study. We then use the best model to analyze the potential differences between the two classes.

C.9 shows that, for all studies, the FMM model had lower BICs than the base model.

After fitting the finite mixture models, we obtain the predicted posterior probabilities for each class. These probabilities indicate the likelihood of each observation (i.e., CV) belonging to a specific class. We choose the one with the highest predicted posterior probability to assign a class to each observation. This approach ensures that each CV is assigned to the class with the highest probability of belonging, maximizing the model's overall fit.

With the CVs assigned to their respective classes, we then compute the correlation coefficients between PC1 and callbacks separately for each class. This allows

us to investigate how the relationship between social perceptions and callbacks differs across job types or classes. By examining these correlations, we can better understand the role of social perceptions in driving callback rates for different types of jobs and determine if specific job characteristics are more or less sensitive to social perceptions.

We then compare the job characteristics between classes across these studies. C.10 presents the correlation coefficients between PC1 and callbacks for each study under three different scenarios: the base model without latent classes, within Class 1, and Class 2. For instance, in the Bertrand study, the correlation between PC1 and callbacks is 0.05 in the base model but increases to 0.57 within Class 1. However, no correlation coefficient is reported for Class 2 in the Bertrand study, as indicated by "NA" in the table. This implies that there is no variation in callbacks within this class—either all CVs received callbacks (all 1s) or none of them did (all 0s).

Similarly, we can observe varying correlation coefficients within each class compared to the base model for the other studies. These results highlight that the relationship between social perceptions and callbacks is not uniform across different job types and that some jobs might exhibit stronger or weaker associations between social perceptions and callbacks than others.

For the following analyses, we will focus only on Farber and Nunley, as these studies produced classes with enough callback variation for us to draw meaningful conclusions. C.11 presents the text data from job titles and descriptions and other relevant variables with the largest difference between classes. This comparison allows us to understand the qualitative differences between the two better.

From the table, we can observe noticeable differences between the two classes regarding job characteristics. In the Nunley dataset, Class 1 has a higher prevalence of Manager, Analyst, Management, Finance, and Specialist positions, while Representative, Insurance, Entry Level, and Sales roles dominate Class 2. This suggests that Class 1 might comprise more advanced, specialized, or managerial positions, whereas Class 2 consists of more entry-level or sales-oriented roles.

In the Farber dataset, Class 1 is characterized by a higher presence of Health Care and Social Assistance, Professional, Scientific, and Technical Services, Finance and Insurance, Retail Trade, and Manufacturing industries. In contrast, Class 2 is more heavily represented in Arts, Entertainment, and Recreation, Educational Services, Repair and Maintenance, and Real Estate and Rental and Leasing industries. Ad-

ditionally, the Administrator occupation is more common in Class 1, while Class 2 shows a more diverse range of industries, focusing on non-managerial positions. This indicates that Class 1 might be associated with more professional or technical industries. In contrast, Class 2 is linked to a broader variety of roles, primarily in service-oriented and less specialized sectors.

With these qualitative differences, we can analyze the differences in the correlation between callbacks and the first principal component (PC1) for each study and class. In the Farber study, the correlation between callbacks and PC1 (representing positive social perceptions, including warmth and competence) is higher in Class 2 (0.26) than in Class 1 (0.17). Class 1 contains job titles associated with more professional and technical roles, while Class 2 comprises a more diverse range of service-oriented and less specialized roles. This suggests that the broader variety of roles in Class 2 might benefit more from positive social perceptions, especially warmth, when it comes to receiving callbacks.

In contrast, the Nunley study shows a slightly higher correlation between callbacks and PC1 in Class 1 (0.80) than in Class 2 (0.77). Class 1 in this study is characterized by advanced, specialized, or managerial positions, while Class 2 includes entry-level or sales-oriented roles. These findings indicate that competence and warmth are highly valued in both classes but may be slightly more important for receiving callbacks in Class 1, which consists of more specialized positions.

However, to compare the correlation coefficients across classes, we need to use the Fisher transformation due to the Nunley study's highly skewed distribution of correlation coefficients. The Fisher transformation is given by:

$$r' = \frac{1}{2} \ln \frac{1+r}{1-r}.$$

After applying the Fisher transformation and calculating the z-values for the difference between correlation coefficients using the formula:

$$z = \frac{r'_1 - r'_2}{\sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}},$$

we find that for the Farber study, $z = 0.200$, and for Nunley, $z = 0.124$.

The magnitudes of the $z$-scores indicate the difference size between the two classes' correlation coefficients. In this case, neither the Farber study's $z$-score (0.200)

nor the Nunley study's $z$-score (0.124) exceed the critical value, indicating that the differences in correlation coefficients between the classes are not statistically significant at the $p < 0.05$ level.

It is important to note that PC1 represents warmth and competence, and this amalgamation of warmth and competence makes it difficult to distinguish the specific role each dimension plays, particularly in entry-level positions where the contribution of warmth and competence to callbacks may differ.

We conducted an exploratory analysis using partial correlations to better understand the interplay between warmth and competence in influencing callbacks for various job positions. Partial correlations allow us to examine the relationship between two variables while controlling for the influence of one or more other variables. We assessed the relationships between warmth and competence scores separately, controlling for the other dimension, as seen in C.12.

These results suggest that different job categories may have distinct relationships between warmth and competence. In Farber's study, warmth appears to be positively related to callbacks in both job classes, while competence is negatively related. However, the magnitude of the relationship between warmth and callbacks is greater in Class 2, which includes jobs requiring more social interaction, aligning with our previous qualitative analysis. The negative relationship between competence and callbacks is stronger in Class 1, suggesting that competence may not be as crucial for jobs that involve more routine tasks and lower technical skills.

In contrast, Nunley's study shows a mixed relationship for warmth, with a negative relationship in Class 1 and a positive one in Class 2 and a positive relationship between competence and callbacks in both classes. The negative relationship between warmth and callbacks in Class 1 is larger in magnitude than the positive relationship in Class 2, indicating that warmth may be less important or even detrimental for jobs requiring higher cognitive and technical skills. Conversely, the positive relationship between competence and callbacks is notably stronger in Class 1, which consists of jobs requiring higher cognitive and technical skills, highlighting the importance of competence in these positions.

It is important to note that this analysis is exploratory, and further research is needed to validate these findings. However, these results do provide preliminary evidence that the relationships between warmth, competence, and job callbacks may be more complex than initially thought.

Fig. C.2.– Fig. C.5. display various measures of heterogeneity of a random effects meta-model, where the effect size is $\rho$(callback, PC1).

## C.3 Supplementary Figures

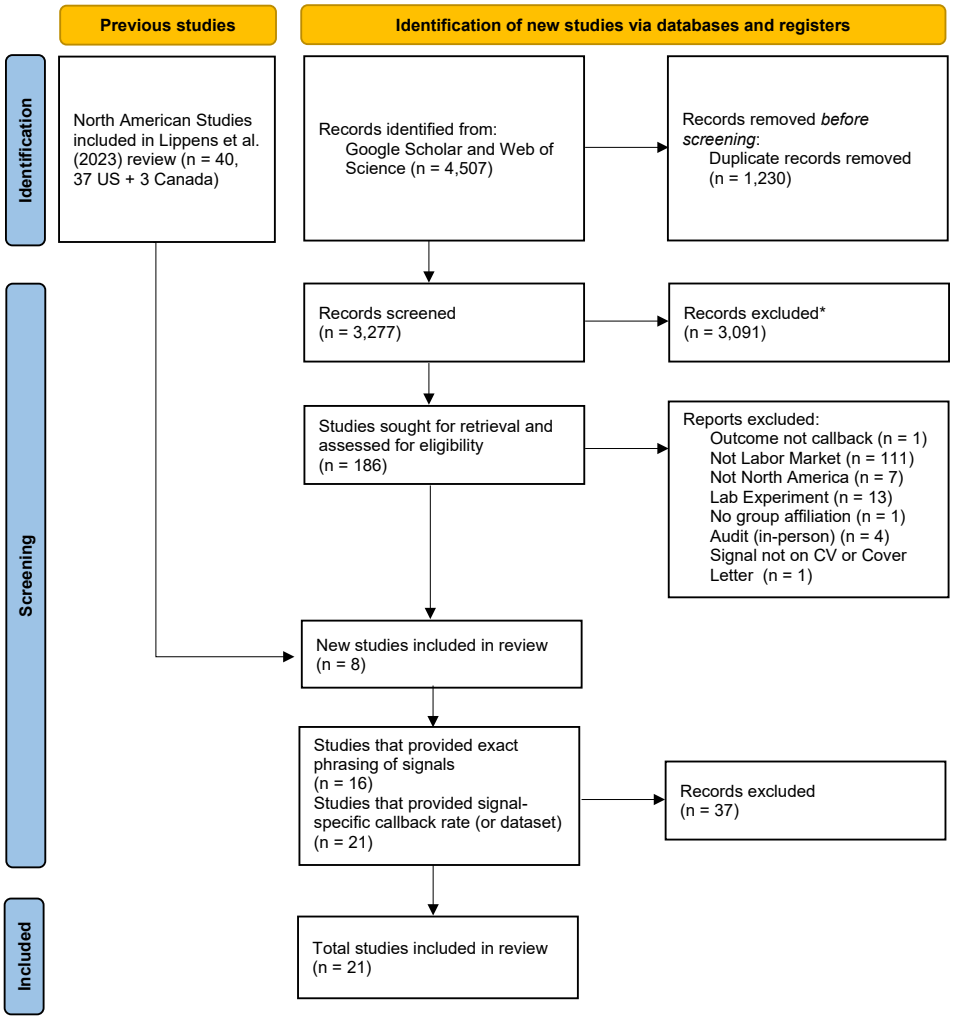



Figure C.1: Prisma chart

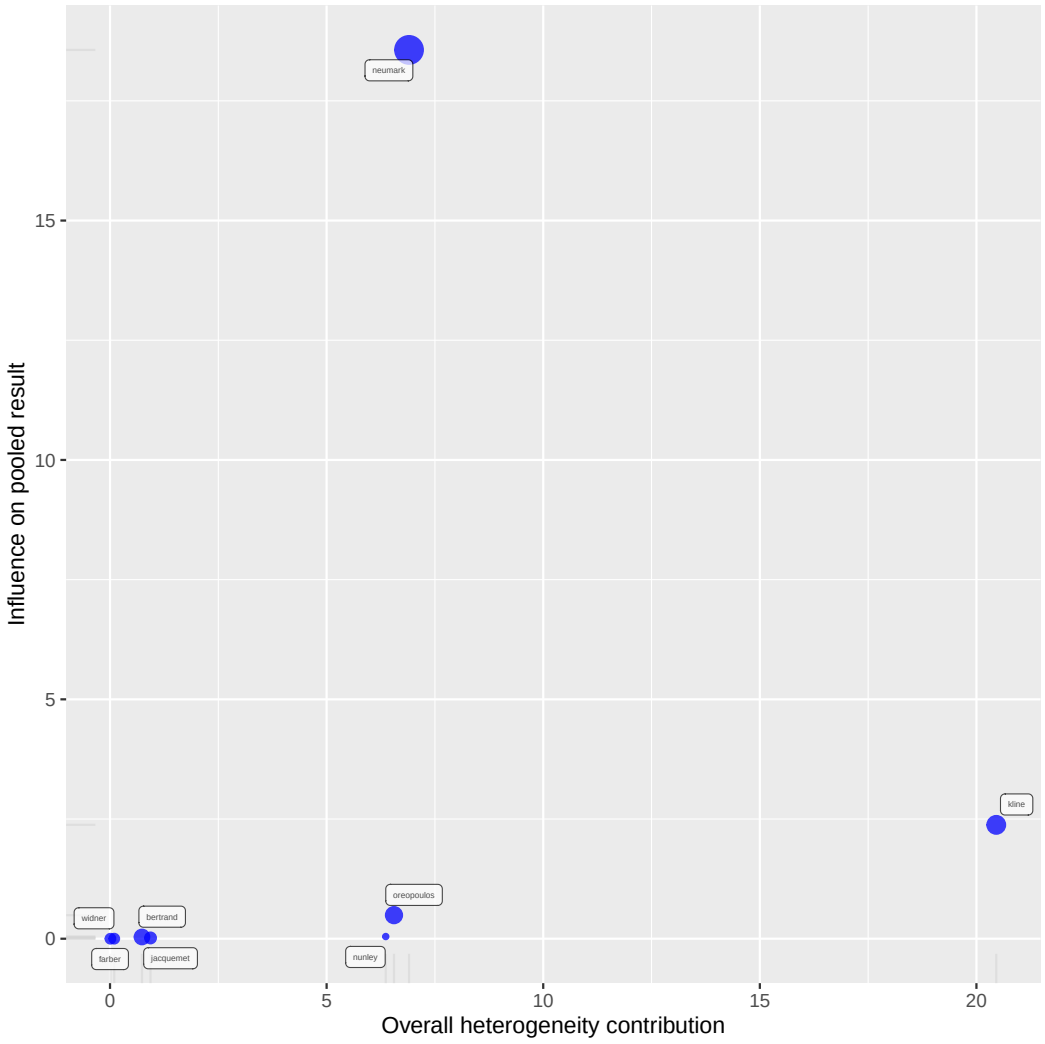



Figure C.2: The Baujat plot

Baujat et al., 2002 is a diagnostic plot used to identify studies that disproportionately contribute to heterogeneity in a meta-analysis. The plot displays the contribution of each study to the overall heterogeneity (measured by Cochran's Q) on the $x$-axis and its impact on the pooled effect size (defined as the standardized squared difference between the overall estimate based on an equal-effects model with and without the $i^th$ study included in the model) on the $y$-axis. Our analysis found that Kline et al., 2022b significantly influenced the heterogeneity but did not significantly affect the pooled effect size. On the other hand, Neumark et al., 2019 contributed moderately to the overall heterogeneity but substantially impacted the pooled effect size.

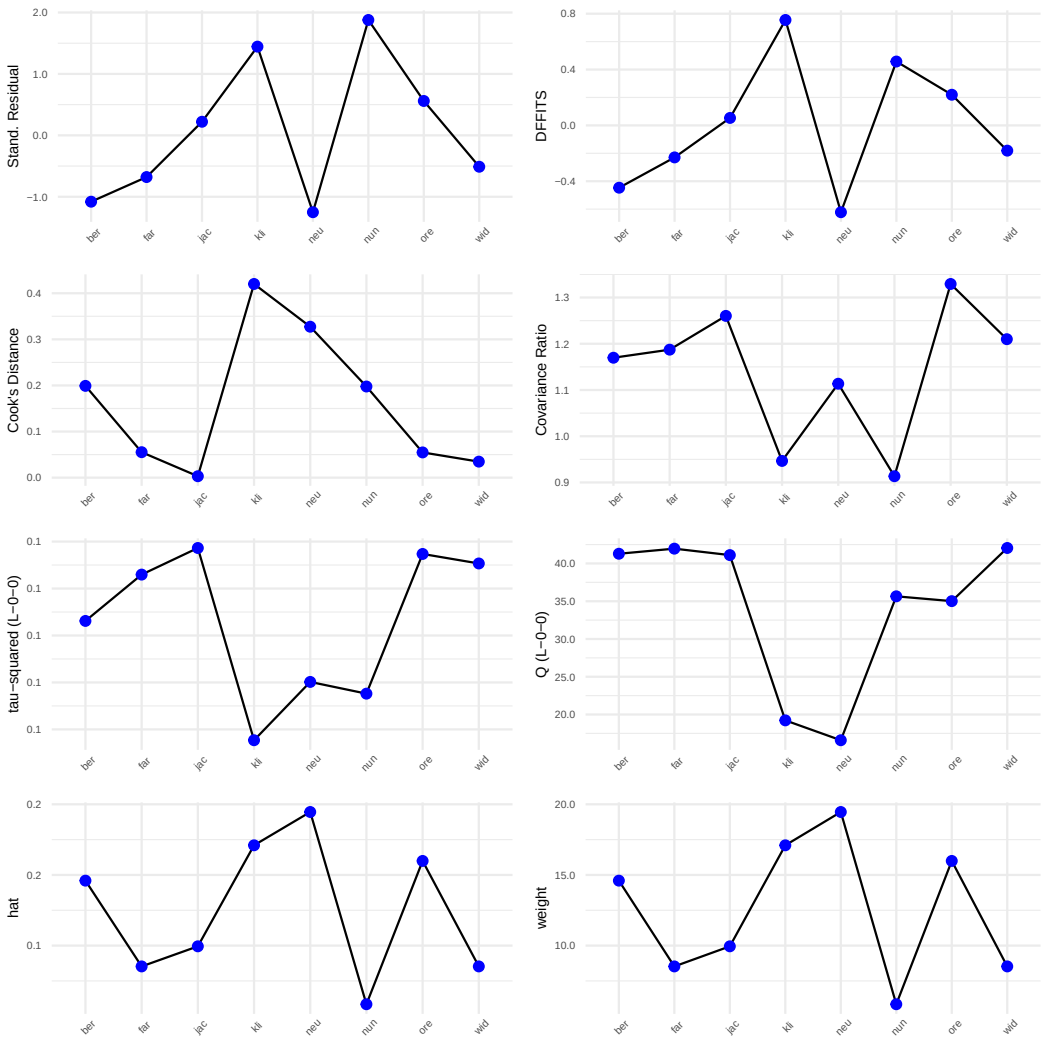



Figure C.3: Different influence measures

The plot displays different influence measures for each study, which help to identify potential outliers that do not fit well into the meta-analysis model. No study was detected as an outlier based on these measures.

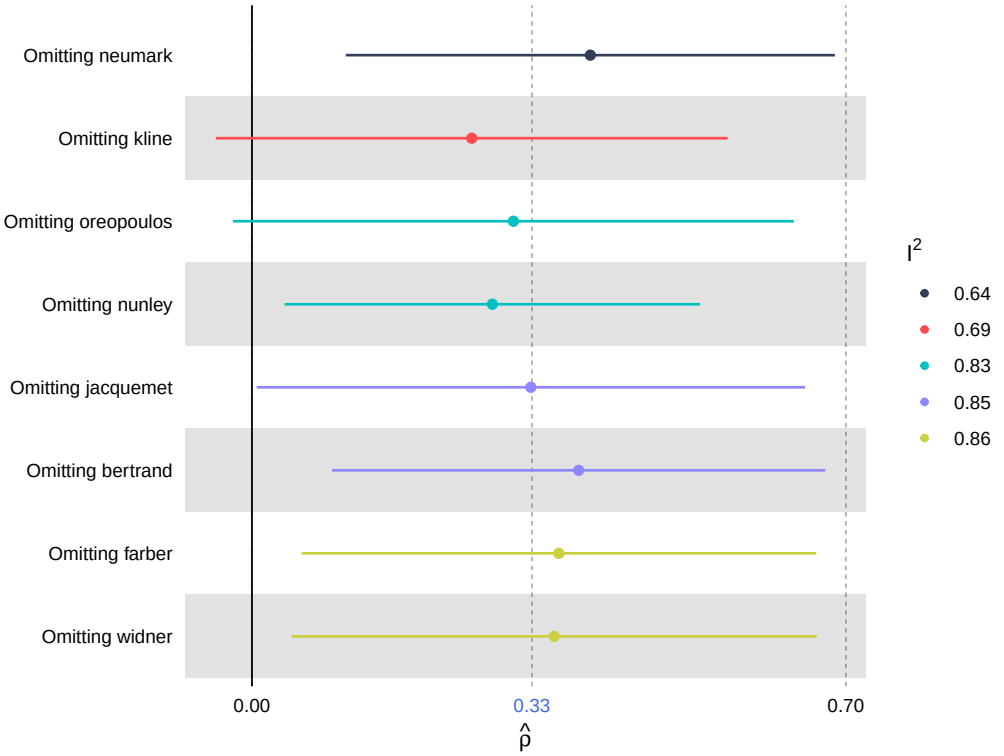



Figure C.4: Heterogeneity effect

The plot displays the overall effect and $I^2$ heterogeneity of all meta-analyses with $\rho$(callback, PC1) as effect size that were conducted using the leave-one-out method. The forest plot is sorted by the $I^2$ value of the leave-one-out meta-analyses. The results show that excluding Neumark et al., 2019 leads to the largest reduction in $I^2$, reducing it from 82% to 64%.

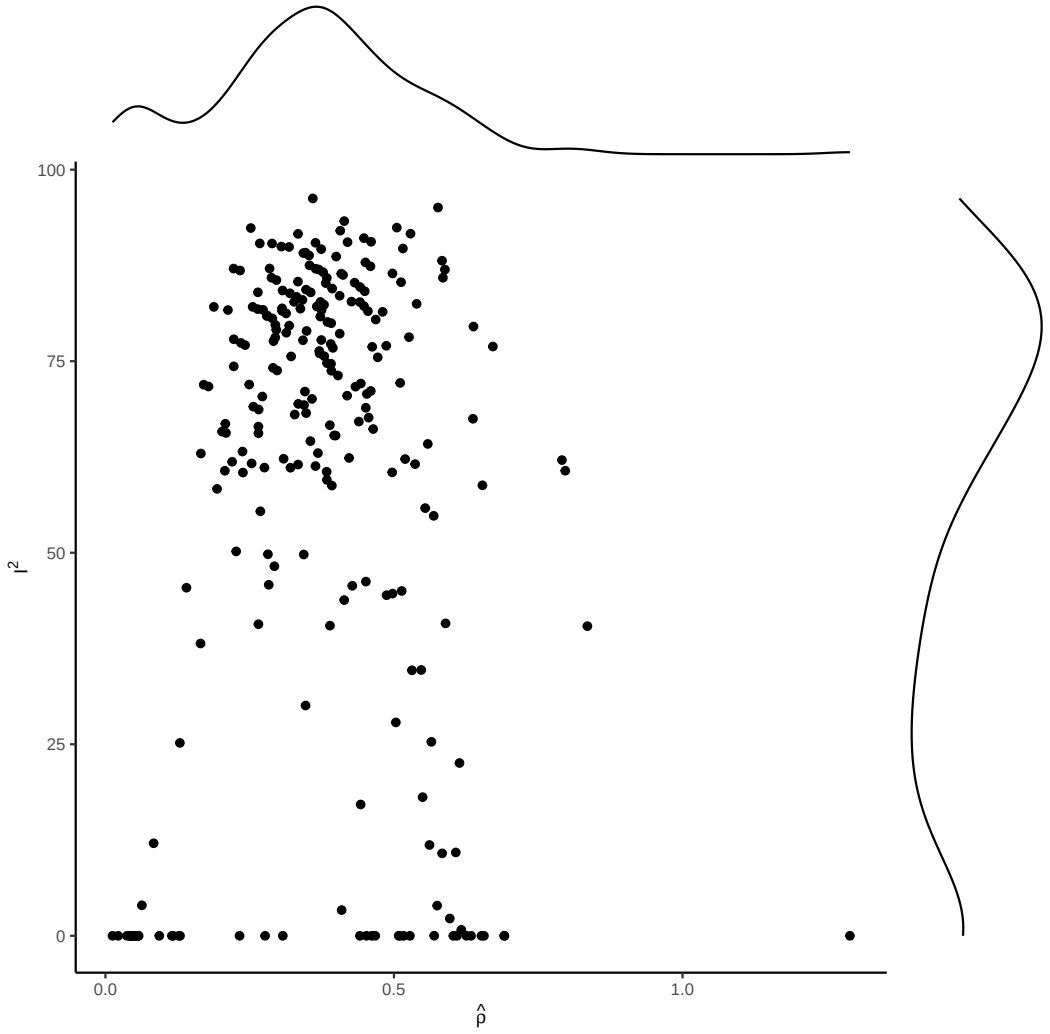



Figure C.5: GOSH plot

We implemented a Graphical display of heterogeneity (GOSH) plot analysis. For this analysis, we fit all possible subsets $2^{k-1}$ of our $k$ included studies. Each subset's pool effect size $\hat{\rho}$ is plotted on the x-axis, and the between-study heterogeneity $I^2$ on the y-axis. Three (k-means, DBSCAN, gmm) clustering algorithms are used to determine patterns in the above scatter plot. The three algorithms did not consistently identify clusters therefore, we conclude that based on this analysis, no single study needs to be excluded from estimating the meta-model.

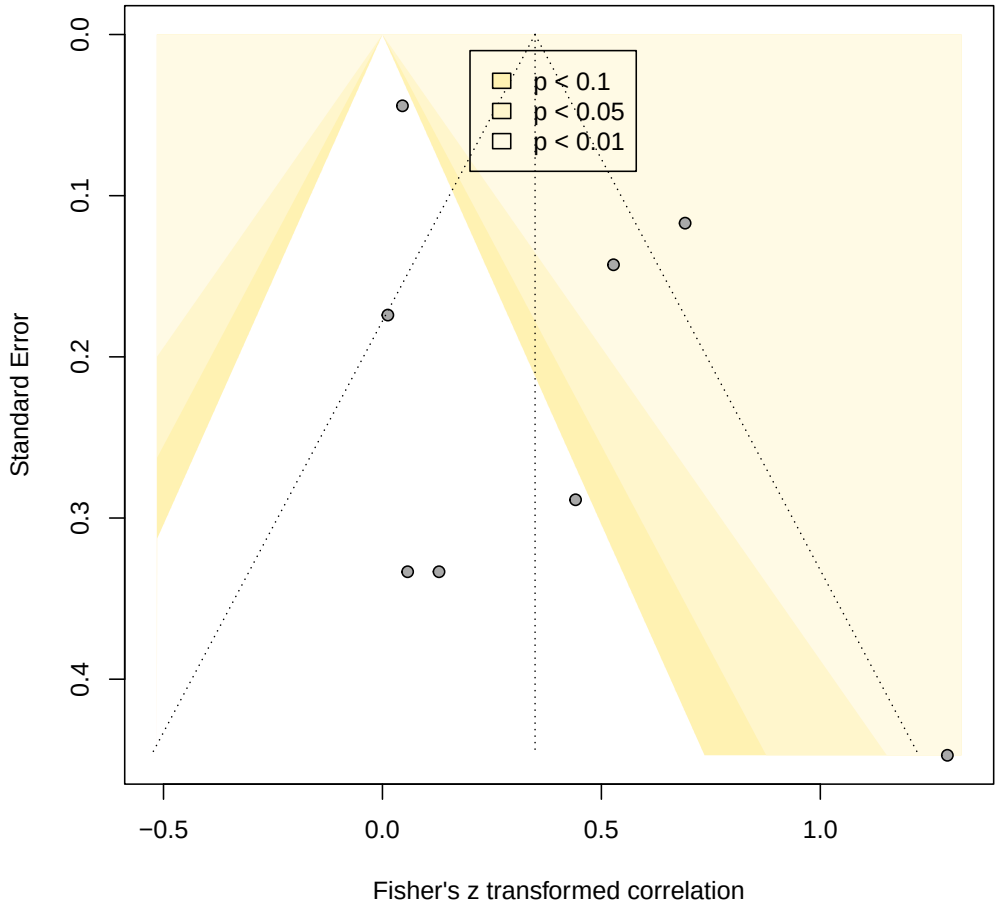



Figure C.6: Funnel plot

The plot shows the effect size of each study (expressed as the standardized mean difference) on the x-axis and the standard error (from large to small) on the y-axis. An idealized funnel shape is included in the plot to facilitate interpretation. The dotted vertical line in the middle of the funnel represents the average effect size. The funnel plot displays three shaded regions, focusing on the $p < 0.05$ and $p < 0.01$ regions, where effect sizes are considered significant. Without small-study effects, our studies should follow the funnel shape delineated in the plot. The funnel plot shows no asymmetrical pattern, which may indicate publication bias in the dataset. Egger's regression test suggests the absence of publication bias in our analysis (intercept = 1.8, 95% CI = -0.25 – 3.85, t = 1.72, p = .14). However, it is essential to note that our study includes only $k = 7$ studies, which may limit the reliability of the test, as it is prone to bias when $k < 10$. It is important to note that our analysis does not directly measure publication bias in the traditional sense. None of the correspondence studies we included examined the correlation between warmth, competence perceptions, and callback rates. Nevertheless, there is a possibility of involuntary or accidental publication bias. For instance, a correspondence study that varies names might only sample names corresponding to higher warmth and competence, resulting in higher callback rates.

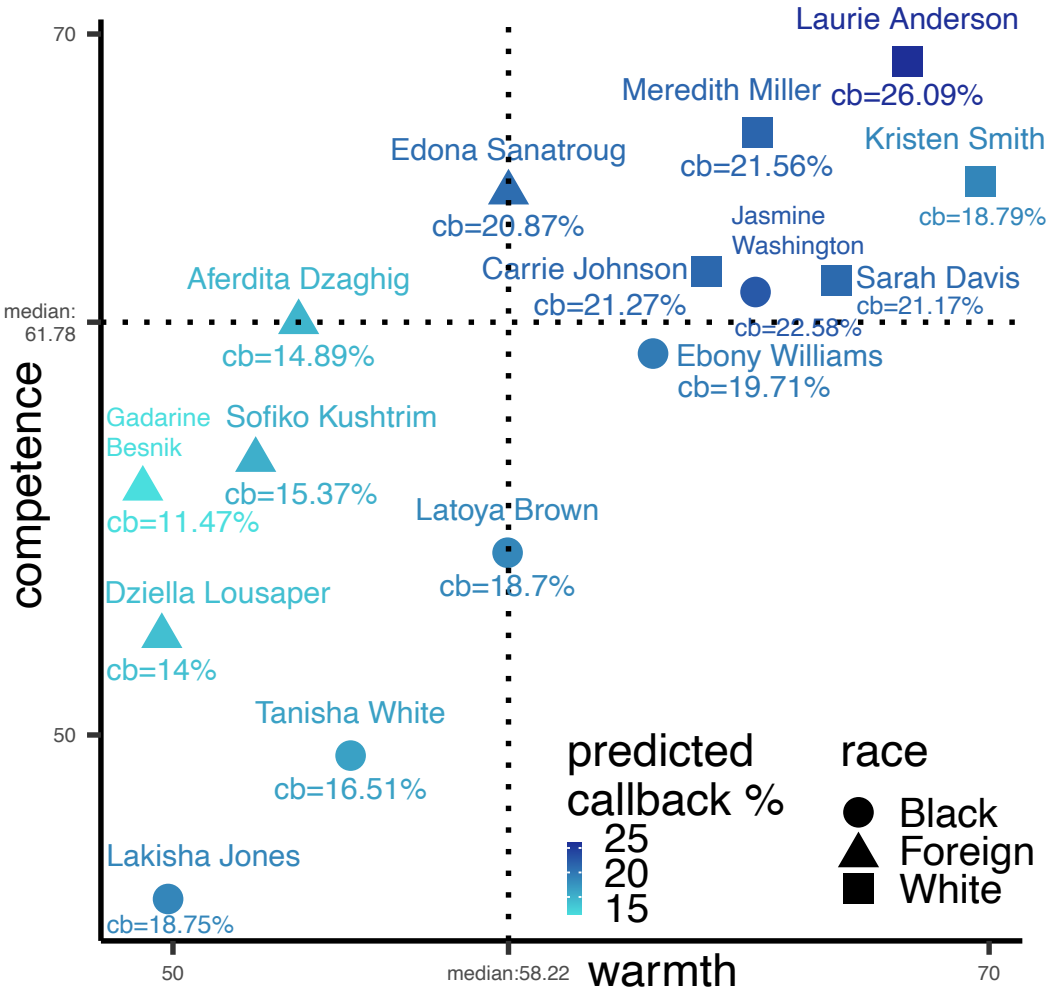



Figure C.7: Warmth and competence ratings

The plot displays warmth and competence ratings for names in Jacquemet and Yannelis, 2012. Each dot's shape represents the name's race, either black, white, or foreign-sounding. The color of the dots corresponds to the predicted callback. The predictions were generated using a linear model of PC1 on callback, with the training set consisting of all names except one. The predicted callback values for each name are displayed beneath the corresponding dot.

## C.4 Supplementary Tables

Table C.1: Published studies for which raw data was obtained. The numbers represent the count of signals (names) per race, gender, and study.

| gender | black | white |
|---|---|---|
| Bertrand and Mullainathan, 2004 | | |
| female | 9 | 9 |
| male | 9 | 9 |
| Farber et al., 2016b | | |
| female | NA | 12 |
| Jacquemet and Yannelis, 2012 | | |
| female | 5 | 5 |
| Kline et al., 2022a | | |
| female | 16 | 17 |
| male | 18 | 17 |
| Neumark et al., 2016 | | |
| female | NA | 313 |
| male | NA | 218 |
| (2017)Nunley et al., 2013 | | |
| female | 2 | 2 |
| male | 2 | 2 |
| Oreopoulos, 2011 | | |
| female | NA | 8 |
| male | NA | 8 |
| Widner and Chicoine, 2011 | | |
| male | NA | 6 |

**Note:** Neumark, Bertrand, Farber, and Kline varied the first name only. Oreopoulos, Flake, Leasure, Widner, and Jacquemet varied the first and the last name.

Table C.2: ICC values for names

Average score intraclass correlations (ICCs) were used as an index of interrater reliability of warmth and competence ratings. A two-way model with random effects for raters and subjects (amount of levels in category) was used. Between-rater agreement was estimated. The unit of analysis was averages. Column "Mean" presents the average warmth and competence ICC score.

| | **Warmth** | | **Competence** | | **Mean** | |
|---|---|---|---|---|---|---|
| | **ICC** | **score** | **ICC** | **score** | **ICC** | **score** |
| jacquemet | 0.98 | excellent | 0.97 | excellent | 0.97 | excellent |
| kline | 0.95 | excellent | 0.96 | excellent | 0.95 | excellent |
| widner | 0.97 | excellent | 0.99 | excellent | 0.98 | excellent |
| bertrand | 0.83 | good | 0.69 | moderate | 0.76 | good |
| neumark | 0.94 | excellent | 0.81 | good | 0.87 | good |
| nunley | 0.65 | moderate | 0.93 | excellent | 0.79 | good |
| oreopoulos | 0.91 | excellent | 0.89 | good | 0.90 | good |
| farber | 0.86 | good | 0.50 | poor | 0.68 | moderate |

Table C.3: ICC values for categories

Average score intraclass correlations (ICCs) were used as an index of interrater reliability of warmth and competence ratings. A two-way model with random effects for raters and subjects (amount of levels in category) was used. Between-rater agreement was estimated. The unit of analysis was averages. Column "Mean" presents the average warmth and competence ICC score.

| | | **Warmth** | | **Competence** | | **Mean** | |
|---|---|---|---|---|---|---|---|
| | **Category** | **ICC** | **Score** | **ICC** | **Score** | **ICC** | **Score** |
| ameri | health | 0.96 | excellent | 0.93 | excellent | 0.95 | excellent |
| hipes | health | 0.96 | excellent | 0.97 | excellent | 0.97 | excellent |
| ishizuka | parenthood | 0.97 | excellent | 0.92 | excellent | 0.95 | excellent |
| namingit | unemployed | 0.96 | excellent | 0.98 | excellent | 0.97 | excellent |
| wrigth | religion | 0.95 | excellent | 0.94 | excellent | 0.95 | excellent |
| yemane | nationality | 0.91 | excellent | 0.93 | excellent | 0.92 | excellent |
| mishel | sexuality | 0.83 | good | 0.94 | excellent | 0.89 | good |
| farber | age | 0.55 | moderate | 0.79 | good | 0.67 | moderate |
| rivera | wealth | 0.82 | good | 0.63 | moderate | 0.72 | moderate |
| tilcsik | sexuality | 0.64 | moderate | 0.51 | moderate | 0.58 | moderate |
| bailey | sexuality | 0.04 | poor | 0.95 | excellent | 0.50 | poor |
| correll | parenthood | 0.87 | good | 0.00 | poor | 0.43 | poor |
| figinski | military | 0.00 | poor | 0.00 | poor | 0.00 | poor |
| kline | sexuality | 0.00 | poor | 0.94 | excellent | 0.47 | poor |
| neumark | age | 0.08 | poor | 0.33 | poor | 0.21 | poor |
| thomas | wealth | 0.00 | poor | 0.97 | excellent | 0.49 | poor |

Table C.4: 95% CI for $\rho$ by study

Random effects model of 10 studies with 418 observations using the inverse variance method. Restricted maximum-likelihood estimator for $\tau^2$ and Hartung-Knapp adjustment (df = 8). Confidence intervals for $\tau^2$ and $\tau$ were estimated using the Q-Profile method. Fisher's z transformation was used for correlations.

| | | 95% CI | | | |
|---|---|---|---|---|---|
| **Study** | $\rho$ | **Lower** | **Upper** | **p-value** | **SE** |
| bertrand | 0.616 | 0.378 | 1.060 | 0.000 | 0.174 |
| farber | 0.408 | -0.220 | 1.086 | 0.194 | 0.333 |
| flake | 0.900 | 1.241 | 1.700 | 0.000 | 0.117 |
| gorzig | 0.826 | 0.585 | 1.767 | 0.000 | 0.302 |
| jacquemet | 0.845 | 0.715 | 1.763 | 0.000 | 0.267 |
| kline | 0.900 | 1.241 | 1.700 | 0.000 | 0.117 |
| neumark | 0.631 | 0.565 | 0.923 | 0.000 | 0.091 |
| nunley | 0.740 | 0.373 | 1.826 | 0.000 | 0.302 |
| oreopoulos | 0.565 | 0.334 | 0.946 | 0.000 | 0.156 |
| widner | 0.915 | 0.904 | 2.210 | 0.000 | 0.333 |
| **Pooled $\rho$** | **0.780** | **0.759** | **1.330** | **0.000** | **0.126** |

Table C.5: 95% CI for $\rho$ by study

Random effects model of 16 studies with 7830 observations using the inverse variance method. Restricted maximum-likelihood estimator for $\tau^2$ and Hartung-Knapp adjustment (df = 15). Confidence intervals for $\tau^2$ and $\tau$ were estimated using the Q-Profile method. Fisher's z transformation was used for correlations.

| | | 95% CI | | | |
|---|---|---|---|---|---|
| **Study** | $\rho$ | **Lower** | **Upper** | **p-value** | **SE** |
| ameri | 0.574 | 0.514 | 0.794 | 0.000 | 0.072 |
| bailey | 0.546 | 0.474 | 0.752 | 0.000 | 0.071 |
| correll | 0.645 | 0.626 | 0.907 | 0.000 | 0.072 |
| farber | 0.602 | 0.616 | 0.778 | 0.000 | 0.069 |
| figinski | 0.342 | 0.216 | 0.496 | 0.000 | 0.072 |
| hipes | 0.699 | 0.724 | 1.142 | 0.000 | 0.070 |
| ishizuka | 0.710 | 0.749 | 1.026 | 0.000 | 0.062 |
| kline | 0.416 | 0.304 | 0.582 | 0.000 | 0.071 |
| mishel | 0.456 | 0.304 | 0.582 | 0.000 | 0.072 |
| namgnit | 0.776 | 0.922 | 1.149 | 0.000 | 0.058 |
| neumark | 0.724 | 0.802 | 1.030 | 0.000 | 0.058 |
| rivera | 0.583 | 0.526 | 0.808 | 0.000 | 0.072 |
| thomas | 0.723 | 0.813 | 0.594 | 0.000 | 0.060 |
| tilcsik | 0.437 | 0.327 | 0.697 | 0.000 | 0.072 |
| wright | 0.696 | 0.789 | 1.070 | 0.000 | 0.070 |
| yemane | 0.639 | 0.724 | 0.791 | 0.000 | 0.017 |
| **Pooled $\rho$** | **0.595** | **0.579** | **0.792** | **0.000** | **0.050** |

Table C.6: Pooling effect sizes competence, warmth, and callback for the categories race and gender

| | | 95% CI | | | |
|---|---|---|---|---|---|
| | **Estimate** | **Lower** | **Upper** | **p-value** | **SE** |
| **Competence**[1] | | | | | |
| black[2] | -11.52 | -23.74 | 0.71 | 0.06 | 3.84 |
| female[3] | -3.07 | -9.56 | 3.42 | 0.32 | 2.91 |
| **Warmth**[1] | | | | | |
| black[2] | -6.72 | -19.19 | 5.76 | 0.19 | 3.92 |
| female[3] | 2.88 | -4.39 | 10.16 | 0.40 | 3.27 |
| **Callback**[4] | | | | | |
| black[5] | 0.79 | -0.51 | 0.04 | 0.07 | 0.09 |
| female[6] | 1.02 | -0.03 | 0.06 | 0.36 | 0.01 |

[1] Statistic is a warmth/competence rating expressed on a scale from 0 to 100. Models involve the inverse variance method and a restricted maximum-likelihood estimator for $\tau^2$. The Q-Profile method was used to compute the confidence interval of $\tau^2$ and $\tau$, and a Hartung-Knapp (HK) adjustment was applied for the random effects model, with degrees of freedom set to 10.

[2] k=4 studies, o=687 observations.

[3] k=11 studies, o=816 observations.

[4] The effect size represents a risk ratio. The Mantel-Haenszel method was used to calculate the overall effect size, with the Paule-Mandel estimator used to estimate the between-study variance (tau^2). A random-effects model was employed with the Hartung-Knapp (HK) adjustment to account for potential bias due to small sample sizes. The model had 1 degree of freedom (df = 1).

[5] k=4 studies, o=89872 observations.

[6] k=4 studies, o=143860 observations.

Table C.7: Estimates of linear models of PC1 on callback by category

We used the available data points to fit a linear model for each category, although many categories only had two data points. Empty cells indicate that the relevant statistics could not be computed. We focus exclusively on the slope. It should be noted that the data from published literature is limited, with only a few studies per category and a few levels per category. Our main goal with this analysis is to provide a preliminary glimpse of the trend. We found a positive association between PC1 and callback for some categories, but we found mixed evidence for others. For example, the sexuality category had two studies with a negative slope and two with a positive slope. NaN indicates that there were not enough data points to estimate the relevant statistics.

| | | **Estimate** | **SE** | **Statistic** | **p-value** |
|---|---|---|---|---|---|
| **wealth** | rivera | -0.84 | | | |
| | thomas | 0.00 | | | |
| **unemployed** | namgnit | 0.02 | 0.01 | 1.74 | 0.33 |
| **sexuality** | bailey | -0.10 | | | |
| | kline | 0.03 | | | |
| | mishel | 0.05 | 0.23 | 0.22 | 0.85 |
| | tilcsik | -0.82 | | | |
| **parenthood** | correll | -0.03 | | | |
| | ishizuka | 0.07 | | | |
| **nationality** | yemane | 0.04 | 0.02 | 1.52 | 0.14 |
| **military** | figinski | 1.41 | | | |
| **health** | ameri | 0.00 | | | |
| | hipes | 0.07 | | | |
| **age** | farber | 0.04 | 0.02 | 1.63 | 0.18 |
| | neumark | 0.20 | 0.23 | 0.87 | 0.54 |
| | wright | 0.01 | 0.00 | 1.68 | 0.14 |

Table C.8: Mixed effects models with varying independent variables

Our findings suggest that PC1 is a positive and significant predictor of callback. The table investigates how PC1 compares as a predictor to categorical variables that are commonly used in correspondence studies. To this end, we fit three mixed-effects models, of different predictors (PC1, race, PC1+race) on callback. Our analysis reveals that the $R^2$ value is highest for model three, as expected. However, we also observe that the $R^2$ value is substantially higher (4.36) for model PC1 compared to the model with race only (.95). Notably, our results show that race is never a significant predictor of callback in either model one or model three, whereas PC1 is a significant predictor in both models. These findings underscore the importance of social perception as a valuable predictor of callback.

| | | | | **95% CI** | | |
|---|---|---|---|---|---|---|
| | $\beta$ | **SE** | **p-value** | **Lower** | **Upper** | $R^2$ |
| **callback ~ race** | | | | | | |
| intrcpt | -2.07 | 0.30 | 0.00 | -2.67 | -1.48 | 0.95 |
| Black | 0.71 | 0.79 | 0.37 | -0.85 | 2.27 | 0.95 |
| **callback ~ PC1** | | | | | | |
| intrcpt | -1.98 | 0.25 | 0.00 | -2.48 | -1.48 | 4.36 |
| PC1 | 0.99 | 0.30 | 0.00 | 0.41 | 1.57 | 4.36 |
| **callback ~ race + PC1** | | | | | | |
| intrcpt | -2.14 | 0.30 | 0.00 | -2.73 | -1.56 | 6.13 |
| PC1 | 1.06 | 0.30 | 0.00 | 0.45 | 1.67 | 6.13 |
| Black | 1.19 | 0.79 | 0.13 | -0.36 | 2.74 | 6.13 |

**Note:** Mixed-Effects Models (k = 644; $\tau^2$ estimator: ML)

Table C.9: Comparison of BIC values for base (logit, no classes) and FMM models in each study

| | **N** | **df** | **BIC** |
|---|---|---|---|
| *Farber* | | | |
| Base | 8899 | 2 | 6227.6 |
| FMM | 8665 | 20 | 6160.1 |
| *Oreopoulos* | | | |
| Base | 12910 | 2 | 8351.3 |
| FMM | 12910 | 26 | 8302.0 |
| *Neumark* | | | |
| Base | 31523 | 2 | 27007.7 |
| FMM | 31523 | 8 | 26172.0 |
| *Nunley* | | | |
| Base | 9396 | 2 | 8463.0 |
| FMM | 9396 | 59 | 7613.8 |
| *Kline* | | | |
| Base | 74946 | 2 | 82759.8 |
| FMM | 68297 | 22 | 74504.6 |
| *Bertrand* | | | |
| Base | 5635 | 2 | 3135.5 |
| FMM | 5635 | 10 | 3118.5 |

Table C.10: Correlation coefficients $r$(PC1, callbacks) in the base model (logit, no classes) and within each class for each study

| | Base (no class) | | Class 1 | | Class 2 | |
|---|---|---|---|---|---|---|
| Study | $r$ (base) | N | $r$ | Proportion (%) | $r$ | Proportion (%) |
| farber | 0.07 | 9,240 | 0.17 | 64.92 | 0.26 | 35.08 |
| nunley | 0.86 | 9,396 | 0.80 | 23.88 | 0.77 | 76.12 |
| bertrand | 0.05 | 5,635 | 0.57 | 98.76 | NA | 1.24 |
| kline | 0.60 | 74,946 | NA | 22.67 | 0.20 | 77.33 |
| oreopoulos | 0.49 | 12,910 | -0.61 | 90.08 | NA | 9.92 |
| neumark | 0.39 | 31,523 | NA | 15.32 | NA | 84.68 |

Table C.11: Mean differences in job characteristics between classes for Nunley et al., 2017 and Farber et al., 2016a

| Variable | Mean (Class 1) | SE (Class 1) | Mean (Class 2) | SE (Class 2) | Z score | p-value |
|---|---|---|---|---|---|---|
| *Nunley* | | | | | | |
| Manager | 0.243 | 0.005 | 0.129 | 0.007 | 11.68 | 1.63E-31 |
| Analyst | 0.104 | 0.004 | 0.000 | 0.000 | 16.22 | 3.82E-59 |
| Management | 0.202 | 0.005 | 0.101 | 0.006 | 11.04 | 2.36E-28 |
| Finance | 0.195 | 0.005 | 0.111 | 0.006 | 9.33 | 1.03E-20 |
| Specialist | 0.078 | 0.003 | 0.003 | 0.001 | 13.16 | 1.45E-39 |
| Representative | 0.097 | 0.004 | 0.225 | 0.009 | -15.94 | 3.46E-57 |
| Insurance | 0.086 | 0.003 | 0.233 | 0.009 | -18.78 | 1.08E-78 |
| Level | 0.005 | 0.001 | 0.200 | 0.008 | -36.23 | 1.9E-287 |
| Entry | 0.003 | 0.001 | 0.201 | 0.008 | -37.43 | 1.2E-306 |
| Sales | 0.194 | 0.005 | 0.562 | 0.010 | -34.14 | 1.9E-255 |
| *Farber* | | | | | | |
| Health Care and Social Assistance | 0.260 | 0.006 | 0.074 | 0.005 | 20.77 | 7.3E-96 |
| Professional, Scientific, and Technical Services | 0.240 | 0.006 | 0.079 | 0.005 | 18.40 | 1.2E-75 |
| Finance and Insurance | 0.099 | 0.004 | 0.009 | 0.002 | 15.84 | 1.72E-56 |
| Retail Trade | 0.103 | 0.004 | 0.037 | 0.003 | 10.79 | 3.89E-27 |
| Manufacturing | 0.069 | 0.003 | 0.003 | 0.001 | 13.85 | 1.2E-43 |
| Administrator | 0.477 | 0.006 | 0.549 | 0.009 | -6.65 | 2.96E-11 |
| Arts, Entertainment, and Recreation | 0.000 | 0.000 | 0.105 | 0.006 | -25.60 | 1.4E-144 |
| Educational Services | 0.004 | 0.001 | 0.113 | 0.006 | -25.00 | 5.9E-138 |
| Repair and Maintenance | 0.000 | 0.000 | 0.175 | 0.007 | -33.37 | 3.4E-244 |
| Real Estate and Rental and Leasing | 0.000 | 0.000 | 0.215 | 0.008 | -37.26 | 6.7E-304 |

Table C.12: Partial correlation.

Partial Correlation coefficients between warmth, competence, and callbacks within each class (i.e., the estimate of the correlation between warmth and callbacks, controlling for other competence (and vice-versa).

| | Partial Correlation | |
|---|---|---|
| | Warmth | Competence |
| *Farber* | | |
| Class 1 | 0.29 | -0.2 |
| Class 2 | 0.36 | -0.16 |
| *Nunley* | | |
| Class 1 | -0.34 | 0.83 |
| Class 2 | 0.28 | 0.47 |

Table C.13: Published studies from which categories were extracted

| Reference | Category |
|---|---|
| Ameri et al., 2018 | Health |
| Bailey et al., 2013 | Sexuality |
| Correll et al., 2007 | Parenthood |
| Farber et al., 2016a | Age |
| Figinski, 2017 | Military |
| Hipes et al., 2016 | Health |
| Ishizuka, 2021 | Parenthood |
| Kline et al., 2022b | Sexuality |
| Mishel, 2016 | Sexuality |
| Namingit et al., 2021 | Unemployed |
| Neumark et al., 2019 | Age |
| Rivera and Tilcsik, 2016 | Wealth |
| Thomas, 2018 | Wealth |
| Tilcsik, 2011 | Sexuality |
| Wright et al., 2013 | Religion |
| Yemane and Fernández-Reino, 2021 | Nationality |

## References


Ameri, M., Schur, L., Adya, M., Bentley, F. S., McKay, P., & Kruse, D. (2018). The disability employment puzzle: A field experiment on employer hiring behavior. *ILR Review*, *71*(2), 329–364.

Bailey, J., Wallace, M., & Wright, B. (2013). Are gay men and lesbians discriminated against when applying for jobs? A four-city, internet-based field experiment. *Journal of Homosexuality*, *60*(6), 873–894.

Baujat, B., Mahé, C., Pignon, J.-P., & Hill, C. (2002). A graphical method for exploring heterogeneity in meta-analyses: Application to a meta-analysis of 65 trials. *Statistics in Medicine*, *21*(18), 2641–2652.

Bertrand, M., & Mullainathan, S. (2004). Replication data for: Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. https://doi.org/10.3886/E116023V1

Cochran, W. G. (1954). Some methods for strengthening the common $\chi$ 2 tests. *Biometrics*, *10*(4), 417–451.

Correll, S. J., Benard, S., & Paik, I. (2007). Getting a job: Is there a motherhood penalty? *American Journal of Sociology*, *112*(5), 1297–1338.

Farber, H. S., Silverman, D., & Von Wachter, T. (2016a). Determinants of callbacks to job applications: An audit study. *American Economic Review*, *106*(5), 314–318.

Farber, H. S., Silverman, D., & Von Wachter, T. (2016b). Replication data for: Determinants of callbacks to job applications: An audit study. https://doi.org/10.3886/E113434V1

Figinski, T. F. (2017). The effect of potential activations on the employment of military reservists: Evidence from a field experiment. *ILR Review*, *70*(4), 1037–1056.

Hartung, J., & Knapp, G. (2001a). On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Statistics in Medicine*, *20*(12), 1771–1782.

Hartung, J., & Knapp, G. (2001b). A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine*, *20*(24), 3875–3889.

Hipes, C., Lucas, J., Phelan, J. C., & White, R. C. (2016). The stigma of mental illness in the labor market. *Social Science Research*, *56*, 16–25.

IntHout, J., Ioannidis, J. P., Rovers, M. M., & Goeman, J. J. (2016). Plea for routinely presenting prediction intervals in meta-analysis. *BMJ open*, *6*(7), e010247.

Ishizuka, P. (2021). The motherhood penalty in context: Assessing discrimination in a polarized labor market. *Demography*, *58*(4), 1275–1300.

Jacquemet, N., & Yannelis, C. (2012). Indiscriminate discrimination: A correspondence test for ethnic homophily in the Chicago labor market. *Labour Economics*, *19*(6), 824–832.

Jurafsky, D., & Martin, J. H. (2023, January). Speech and language processing (3rd ed. draft). https://web.stanford.edu/~jurafsky/slp3/

Kline, P., Rose, E. K., & Walters, C. R. (2022a). Replication data for: 'systemic discrimination among large U.S. employers'. https://doi.org/10.7910/DVN/HLO4XC

Kline, P., Rose, E. K., & Walters, C. R. (2022b). Systemic discrimination among large us employers. *The Quarterly Journal of Economics*, *137*(4), 1963–2036.

Liljequist, D., Elfving, B., & Skavberg Roaldsen, K. (2019). Intraclass correlation–A discussion and demonstration of basic features. *PloS one*, *14*(7), e0219854.

Mishel, E. (2016). Discrimination against queer women in the U.S. workforce: A résumé audit study. *Socius*, *2*, 2378023115621316.

Namingit, S., Blankenau, W., & Schwab, B. (2021). Sick and tell: A field experiment analyzing the effects of an illness-related employment gap on the callback rate. *Journal of Economic Behavior & Organization*, *185*, 865–882.

Neumark, D., Burn, I., & Button, P. (2016). Replication data for: Experimental age discrimination evidence and the Heckman critique. https://doi.org/10.3886/E113433V1

Neumark, D., Burn, I., & Button, P. (2019). Is it harder for older workers to find jobs? new and improved evidence from a field experiment. *Journal of Political Economy*, *127*(2), 922–970.

Nunley, J. M., Pugh, A., Romero, N., & Seals, A. (2013). Replication data for: Racial discrimination in the labor market for recent college graduates: Evidence from a field experiment (unpublished raw data).

Nunley, J. M., Pugh, A., Romero, N., & Seals, R. A. (2017). The effects of unemployment and underemployment on employment opportunities: Results from a correspondence audit of the labor market for college graduates. *Ilr review*, *70*(3), 642–669.

Oreopoulos, P. (2011). Replication data for: Why do skilled immigrants struggle in the labor market? A field experiment with thirteen thousand resumes. https://doi.org/10.3886/E114770V1

Rivera, L. A., & Tilcsik, A. (2016). Class advantage, commitment penalty: The gendered effect of social class signals in an elite labor market. *American Sociological Review*, *81*(6), 1097–1131.

Sidik, K., & Jonkman, J. N. (2002). A simple confidence interval for meta-analysis. *Statistics in Medicine*, *21*(21), 3153–3159.

Thomas, E. V. (2018). "Why even bother; they are not going to do it?" The structural roots of racism and discrimination in lactation care. *Qualitative Health Research*, *28*(7), 1050–1064.

Thompson, S. G., & Higgins, J. P. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine*, *21*(11), 1559–1573.

Tilcsik, A. (2011). Pride and prejudice: Employment discrimination against openly gay men in the United States. *American Journal of Sociology*, *117*(2), 586–626.

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P., Langan, D., & Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research synthesis methods*, *7*(1), 55–79.

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, *30*(3), 261–293.

Widner, D., & Chicoine, S. (2011). It's all in the name: Employment discrimination against Arab Americans. *Sociological Forum*, *26*(4), 806–823.

Wright, B. R., Wallace, M., Bailey, J., & Hyde, A. (2013). Religious affiliation and hiring discrimination in New England: A field experiment. *Research in Social Stratification and Mobility*, *34*, 111–126.

Yemane, R., & Fernández-Reino, M. (2021). Latinos in the United States and in Spain: The impact of ethnic group stereotypes on labour market outcomes. *Journal of Ethnic and Migration Studies*, *47*(6), 1240–1260.