The Streetlight Effect in Data-Driven Exploration *

Johannes Hoelzemann
University of Vienna

Gustavo Manso
UC Berkeley

Abhishek Nagaraj
UC Berkeley & NBER

Matteo Tranchero
The Wharton School

May 17, 2025

Abstract

We study exploration under uncertainty and show how access to data on past attempts can paradoxically hinder breakthrough discovery. We develop a model of the "streetlight effect" demonstrating that when data highlights attractive but ultimately suboptimal projects, it can narrow exploration and suppress innovation. In a laboratory experiment, we find that revealing the value of an enticing project lowers payoffs and reduces breakthrough discoveries. This drop stems from increased free-riding behavior, which crowds out the generation of new data. We validate our theory in the context of scientific research into the genetic origins of human diseases. To identify the causal impact of past data, we use an instrumental variable that leverages exogenous genetic overlaps between humans and laboratory mice, which reduces research costs for specific genes and leads to prioritized data collection about them. We find that diseases with early evidence of promising genetic targets are 16 percentage points less likely to yield breakthroughs than those where early efforts failed. While competition attenuates the streetlight effect, it does not eliminate it. Our paper provides the first systematic analysis of this phenomenon, outlining the conditions under which data leads agents to look under the lamppost rather than engage in socially beneficial exploration.

JEL Classification: C73, C92, D81, D83.
Keywords: Streetlight effect; Data; Exploration and exploitation; Genetics research; Dynamic public-good problem; Laboratory experiment.

*The experiment was conducted online via the Vienna Center for Experimental Economics in December 2024. The experiment received approval from the University of Vienna and the University of Toronto Research Ethics Board (approval #00038482), and was pre-registered at `https://osf.io/zs2mu/`. We are grateful to Gary Biglaiser, Kevin Bryan, Ben Greiner, Soeren Harrs, Emeric Henry, Ryan Hill, Riitta Katila, Moritz Loewenfeld, Ramana Nanda, Jean-Robert Tyran and participants at the SIE workshop, the MOM workshop at HBS, the MAD conference at Columbia, the Paris Bounded Rationality Workshop at PSE, the SMS Special Conference at Bocconi, the Strategy Science conference, as well as seminars at BU, CEU, GeorgiaTech, Masaryk, Michigan, Minnesota, NBER, Purdue, Rotman, Berkeley, UCLA, Vienna, WashU, Wharton, and WU Vienna for their suggestions. We thank Adit Jain for his outstanding help in designing the experimental platform. Eva Chang and especially Cecil-Francis Brenninkmeijer provided excellent research assistance. Authors are listed in alphabetical order. Corresponding author: `manso@berkeley.edu`.

1 Introduction

A central challenge in medical research is identifying the genetic drivers of human disease from over 19,000 potential gene candidates. Puzzlingly, more than two decades after the Human Genome Project mapped all human genes, the genetic landscape remains relatively underexplored (Edwards et al., 2011). Fewer than 10% of genes have been targeted by approved drugs, despite recognition that many less-studied genes may offer better therapeutic opportunities (Stoeger et al., 2018; Gates et al., 2021). Similar patterns appear in other domains, such as venture capital and industrial R&D, where agents should have strong incentives to search broadly, and yet, collective exploration appears limited and potentially valuable options remain neglected. While this underexploration has been noted in policy discussions, it remains under-theorized in formal economic terms. Understanding its drivers is critical in an era of apparent diminishing returns to research effort (Gordon, 2016; Bloom et al., 2020). Whether these trends reflect intrinsic limits to innovation or a narrowly focused search shaped by innovators' incentive structures remains an open question.

To shed light on this issue, we start by observing that innovative search rarely begins on a blank slate. For instance, a scientist studying a disease typically draws on data from past experiments before selecting a genetic target. We develop a framework to understand how such data shapes the direction of future exploration. Our thinking is motivated by the parable of the *streetlight effect*, where agents disproportionately focus their search in areas with readily available data rather than allocating effort based on scientific theory, market potential, or policy relevance. In our simple model, we show how information on past discoveries can narrow search and, paradoxically, reduce both individual and social returns. This runs counter to the conventional view that accurate data always improves outcomes by reducing uncertainty and making exploration more efficient. Our paper reconciles these perspectives by studying how the streetlight effect can emerge in exploratory search among rational agents and identifying the conditions under which greater data availability may hinder rather than help innovation.

In our strategic multi-armed bandit model, agents choose among risky projects over two periods. Projects can be of low, medium, or high value, but their quality is revealed only through exploration. In each period, the decision-maker uses existing information to choose between investing in a previously explored project or taking a risk by exploring a new one. Exploration costs are borne privately, but the resulting data become publicly available. Within this setup, we examine how providing data on the value of one opportunity influences exploration choices. Our central result is that the impact of

data hinges on the type of project illuminated: information about a medium-value project can *reduce* both individual and group payoffs relative to having data on a low-value project or even no data at all.

The intuition behind our result is that when the medium-value project exceeds the expected return from exploring riskier alternatives, it becomes individually rational for agents to pursue the option highlighted by the data. Since this logic applies to all agents, it induces herding behavior: data reduces uncertainty but also narrows the direction of follow-on investment, collectively suppressing exploration that would result in new data generation. As a result, even rational agents may underexplore due to free riding on the informational externalities of others. Our baseline model, where followers receive the same payoff as initial innovators, reflects settings like science and technology, where knowledge is often non-excludable (Aghion et al., 2008; Hill and Stein, 2025b; Krieger, 2021). When we introduce competition by reducing the rewards for follow-on innovators, the effect persists under moderate rivalry but weakens significantly as competition intensifies. Thus, while competitive pressures can undermine innovation quality through racing dynamics (Hill and Stein, 2025a), their absence may discourage exploration altogether due to the streetlight effect.

Next, we implement an online laboratory experiment to test whether our theoretical predictions hold with human participants. Groups of players take part in a two-period game involving strategic exploration. In the baseline condition, players sequentially select from five unknown options randomly drawn from a known payoff distribution. In the first period, they choose one project without immediate feedback; in the second, they observe all first-round payoffs before selecting again. Payoffs are non-rival and cumulative. We then run the same game but provide players with information on one project—either low, medium, or high in value. The results align with our theory: revealing data on the medium-value project reduces group payoffs by 5% and the likelihood of finding the best outcome by 56%, relative to the no-data baseline. Information on low-value projects has no significant effects, while data on high-value projects improve outcomes. We also vary the degree of payoff rivalry and find that the streetlight effect persists under moderate rivalry but diminishes in magnitude.

While our theoretical and laboratory results are intriguing, it is unclear how extensively the streetlight effect shapes real-world innovation. Therefore, we return to our opening example of scientists looking for disease-related genes to employ as drug targets. Searching for the genetic roots of human diseases closely mirrors our theoretical setup: researchers face over 19,000 protein-coding genes, and pinpointing the right targets involves individually risky exploration that can yield large payoffs for drug development. It is also a collective endeavor, with scientists learning from one another and drawing on

data from published findings. For instance, consider Tangier disease, a rare condition characterized by extremely low levels of HDL cholesterol in the blood. Decades of research had focused on genes that early data suggested as moderately promising, but unlikely to lead to therapeutic breakthroughs—until a scientist definitely linked the disease to mutations in the ABCA1 gene. We leverage these parallels to examine whether dynamics akin to the streetlight effect might steer scientists away from breakthrough discoveries.

We leverage data from DisGeNET, a bibliographic database that links scientific publications to the specific diseases and genes they investigate. Each gene-disease combination is assigned a normalized score reflecting the strength of the supporting scientific evidence, which we use to classify associations as low, medium, or high in scientific value. Our dataset covers genetic discoveries for 3,864 diseases between 1980 and 2019. We use this data to examine how the scientific promise of early discoveries—specifically those made before 2000—shapes subsequent innovation at the disease level. The main analysis cross-sectionally explores the implications of our model using careful controls for disease type and total research effort received. Since the distribution of past data is unlikely to be random, we also employ a complementary identification strategy. We use an instrumental variables approach that exploits variation in genetic similarity between human and mouse genes. Research on a human gene is less costly when scientists can study the same gene in laboratory mice, so genes shared across species tend to be explored earlier (Stoeger et al., 2018). However, diseases differ in the likelihood that such shared genes are of high scientific value. This variation creates quasi-random differences in early data, which we use to instrument for the promise of initial discovery and estimate causal effects.

Our results show that disease areas with promising but suboptimal genes discovered prior to 2000 are 16 percentage points less likely to report a major breakthrough afterward, compared to diseases where all earlier data unveiled low promise targets. In practical terms, discovering a medium-value genetic target delays a breakthrough by an average of 2.8 years, roughly 14% longer than the sample mean of 20.2 years. These findings are confirmed by our instrumental variable framework. Event study estimates show a sharp decline in the number of new genes explored following a medium-value discovery, with no evidence of pre-trends. Consistent with our theory, the mechanism seems to be that early medium-value discoveries reduce the diversity of follow-on research, narrowing exploration and lowering the likelihood of identifying high-impact gene-disease associations. Also in line with our model, we find that the streetlight effect is muted in disease areas with greater competition. Taken together, the empirical evidence from the context of genetic research offers striking support for our theoretical predictions.

Our three-part study contributes to several strands of research. First, we add to a growing literature on how data is generated and how it shapes economic outcomes (Bergemann and Bonatti, 2019; Bessen et al., 2022; Farboodi and Veldkamp, 2020; Jones and Tonetti, 2020). Rather than treating data as a homogeneous commodity, we show that the nature of the data itself (specifically what it illuminates or omits) shapes agents' exploration choices. While our findings intersect with work on data as a public good (e.g., Nagaraj and Tranchero, 2024), they also speak more broadly to information that provides signals on the value of uncertain projects. Notably, our results emerge in a context where we operationalized data as instrumental information, i.e., unbiased and directly payoff-relevant. Our results could be even starker if data were imprecise, biased, or uninformative (Henrich et al., 2010; Cao et al., 2024) or if agents' attention is drawn to salient payoffs (Bordalo et al., 2012, 2013, 2020). Beyond this, we propose a novel mechanism by which data can hinder exploration: by leading agents to implicitly coordinate on certain but dominated projects, thus crowding out new data generation to the detriment of collective outcomes.

Second, we contribute to the literature on strategic experimentation and social learning (Bolton and Harris, 1999; Keller et al., 2005; Klein and Rady, 2011; Hörner et al., 2022). We build on recent experimental work examining behavior under strategic interdependence and informational externalities (Boyce et al., 2016; Hoelzemann and Klein, 2021, 2025). Relative to the commonly studied single-agent bandit problem (Bergemann and Valimaki, 2008), we show how informational spillovers in collective experimentation can create free-rider problems that endogenously limit aggregate data generation and dynamically lower payoffs.[1] We further demonstrate that this mechanism aligns with empirical patterns in scientific research on disease-causing genes (Gates et al., 2021; Edwards et al., 2011; Haynes et al., 2018), illustrating how our framework helps explain real-world search dynamics.

Finally, we contribute to the innovation search literature that examines what drives risky exploration among innovators (March, 1991; Levinthal, 1997; Manso, 2011; Azoulay et al., 2011; Ederer and Manso, 2013; Henry et al., 2022; Arora et al., 2025). We highlight the role of the information environment in driving underexploration. We also build on research exploring how different types of data influence experimentation under technological uncertainty (Ewens et al., 2018; Krieger, 2021). In particular, we show how data might have counterintuitive effects in search, offering a less sanguine outlook for how innovation will be shaped in the age of big data and AI (Agrawal et al., 2024;

[1] A related literature in computer science examines rule-based bandit learning, where a single decision-maker follows fixed decision rules (Vermorel and Mohri, 2005). In contrast, the welfare losses we document arise from incentive misalignment between individually and socially optimal behavior, rather than from bounded rationality.

Cockburn et al., 2019; Kim, 2023). Our evidence on disease-relevant genetic discovery adds to prior work examining how databases shape scientific productivity in the biomedical field (Kao, 2024; Williams, 2013; Tranchero, 2025).

The remainder of the paper proceeds as follows. Section 2 provides an overview of the theoretical framework. Section 3 describes the laboratory experiment. Sections 4 and 5 present the empirical analysis in the context of genetic research. Section 6 concludes.

2 Theoretical Framework

Setup. There are N agents engaged in a search to maximize their individual payoffs, choosing from A projects of initially unknown value, with $N \geq A$. Project payoffs are independent and fall into one of three categories: with probability p_L, a project yields a low payoff (L); with probability p_M, a medium payoff (M); and with probability p_H, a high payoff (H), where $0 \leq L < M < H$ and $p_L + p_M + p_H = 1$. While agents know this distribution in advance, they have no prior information about the specific payoff of any given project. Each agent lives for two periods, is risk-neutral, and discounts future payoffs at zero. Agents cannot communicate directly. This setup reflects real-world environments in which individuals face a set of unknown opportunities, where valuable projects are rare but highly rewarding (Kerr et al., 2014; Manso, 2016).

Dynamics. In each period, the N agents choose projects sequentially in a random order. While they can observe the choices made by earlier movers, they do not yet see the payoffs associated with those choices. Once all agents have selected a project, the payoffs of the chosen projects are revealed to everyone, marking the end of period 1. In period 2, the process repeats with the same order. This time, agents know the payoffs of previously explored projects and can choose either a known project or an unexplored one, whose payoff will again be revealed at the end of the period. Payoffs are cumulative across the two periods, so agents earn the sum of the values of the projects they choose. Importantly, payoffs are non-rival, so if multiple agents select the same project, each receives its full value. Unlike classic payoff externalities in public goods problems, here an agent is affected by others only through the data their choices produce over time (Hoelzemann and Klein, 2021, 2025). This setup mimics competitive markets where organizations conduct parallel R&D. Although projects do not directly compete, the information they generate is valuable to all participants (Krieger, 2021).

Equilibrium without Data. We begin by considering the equilibrium in a setting where no data about project payoffs is available before the game begins. At the start of period 1, all projects offer

the same expected payoff based on the known probability distribution. The sequential structure of the game leads agents to choose different projects to generate more data that can guide decisions in period 2. Since $N \geq A$, agents can implicitly coordinate to explore all projects, so the highest payoff is identified before the second period begins. This means that each agent earns the expected value of a random draw in period 1, followed by the highest available payoff in period 2. The probability that the best discovered project has payoff L is p_L^A, payoff M is $(1-p_H)^A - p_L^A$, and payoff H is $1-(1-p_H)^A$. The collective expected payoff and the likelihood of discovering a high-value project are as follows:

$$\text{[Group Payoff]} \quad N[(p_L + p_L^A)L + (p_M + (1-p_H)^A - p_L^A)M + (p_H + 1 - (1-p_H)^A)H] \tag{1}$$

$$\text{[Group Breakthrough]} \quad 1-(1-p_H)^A \tag{2}$$

Equilibrium with Data on L or H Projects. We now compare the setup above to a scenario where the payoff of one project is publicly revealed at the start of the game. The effect of this data depends on the value of the disclosed project. If the revealed project has a payoff of H, all agents immediately coordinate on it, each earning $2H$, and the group achieves the maximum total payoff of $2H \cdot N$. The probability of a breakthrough is 1, showing how data can lead directly to the best possible outcome by eliminating uncertainty. If, instead, the revealed project has a payoff of L, agents simply avoid that option, and they are back to the original setup with one fewer low-value project. In this case, the group's expected payoff is $N[(p_L + p_L^{(A-1)})L + (p_M + (1-p_H)^{(A-1)} - p_L^{(A-1)})M + (p_H + 1 - (1-p_H)^{(A-1)})H]$, and the probability of a breakthrough is $1-(1-p_H)^{(A-1)}$. These outcomes are similar to the no-data case and converge to it as $A \to \infty$. In other words, when a low-payoff project is revealed and the search space is large enough, there is still dispersed exploration.

Equilibrium with Data on M project. What is arguably more interesting, and so far understudied, is the intermediate case where a medium-value project is revealed. Here, a non-empty parameter space exists in which data can be detrimental due to the streetlight effect. This arises when the payoff from choosing M is attractive enough relative to exploring other, unknown-value projects. If the loss from exploration, given by $M - (p_L L + p_M M + p_H H)$, exceeds the potential gain from exploration, $p_H(H - M)$, then all agents choose the medium project in equilibrium.[2] This leads to the following condition:

[2]Suppose there was an equilibrium where some agents selected other projects. By backward induction, the last such agent would strictly prefer the medium project under this condition.

Assumption 1 ("Medium Project is Good Enough")**.**

$$M > \frac{p_L L + p_H 2H}{2 - p_L - 2p_M} \tag{3}$$

Assumption 1 ensures that selecting the medium project dominates searching for a high-value one. Rational agents choose it in both periods, yielding an expected group payoff of $2M \cdot N$. However, when M is not too large relative to L and H, we can show—perhaps counterintuitively—that payoffs with data are actually lower than those with no data. More formally, we introduce:

Proposition 1 ("Group Payoff with Data on Medium Project")**.** *Under Assumption 1 and if*

$$M < \frac{(p_L + p_L^A)L + (p_H + 1 - (1 - p_H)^A)H}{2 - (1 - p_H)^A + p_L^A - p_M} \tag{4}$$

the group payoff without data is higher than when a medium project is revealed.

Proof. We need to show that the expected group payoff without data exceeds the expected group payoff whenever an M project is revealed upfront. This is true if $N[(p_L + p_L^A)L + (p_M + (1 - p_H)^A - p_L^A)M + (p_H + 1 - (1 - p_H)^A)H] > 2M \cdot N$, which is equivalent to the condition in the proposition. ■

The fact that the medium option offers high individual payoffs does not guarantee it is socially optimal. On the contrary, it can lure agents into avoiding exploration. The known option is tempting when the individual odds of finding the high-value project are low, but at the cost of hurting collective welfare. What rational agents fail to account for are the data externalities created by their own experimentation, even when unsuccessful. This leads to the following two results:

Proposition 2 ("Exploration with Data on Medium Project")**.** *If $\mu|i$ is defined as the unmapped share of projects chosen in period 1 given data i, then under Assumption 1, the following weak inequalities hold: $\mu|H \leq \mu|M \leq \mu|L \leq \mu|\emptyset$*

Proof. The proof directly derives from our preceding discussion. If H is revealed, agents will choose that project, so $\mu = 0$. If M is appealing enough, agents forfeit exploration and only choose the revealed project, so $\mu = 0$. If no data is provided or L is revealed, then agents explore all remaining unknown options in period 1, so $\mu = 1$. ■

Proposition 3 ("Breakthrough with Data on Medium Project")**.** *If $P(H|i)$ is defined as the conditional probability of discovering H given data i, then under Assumption 1, the following strict inequality holds: $P(H|M) < P(H|i)$ where $i \in \{\emptyset, L, H\}$*

Proof. If M is appealing enough, agents never achieve a breakthrough, i.e., never discover H, so $P(H|M) = 0$. If no data is provided or L is revealed ex ante, then agents explore all remaining unknown options in period 1, and the probability that H is discovered at all is $(1 - (1 - p_H)^A)$ and

$(1-(1-p_H)^{(A-1)})$ respectively, which are both strictly greater than 0. The statement is trivially true whenever H is revealed. ■

The streetlight effect arises when the medium payoff is tempting enough for the individual, yet exploration still holds social value—that is, when Assumption 1 and Equation (4) both hold. This requires a skewed payoff distribution. If the distribution of payoffs was symmetric, the expected value of an unknown draw would equal M, making exploration risk-free with a potential upside of $p_H(H-M)$. In that case, Assumption 1 would be violated, and the streetlight equilibrium would not emerge. However, the effect also vanishes under extreme payoff skewness. If the breakthrough is too rare (very small p_H), the expected social value of exploration falls below M, violating Equation (4). If the breakthrough is too common (very large p_H), the private upside $p_H(H-M)$ becomes very attractive, breaking Assumption 1. Thus, the streetlight equilibrium appears only under moderate skewness of the payoff distribution.[3]

The Role of Competition. Our theoretical framework assumes non-rivalry in payoffs, meaning that agents still earn the full reward even if they were not the first to choose a project. While a simplification, this assumption fits reasonably well in fields like scientific research. For instance, Hill and Stein (2025b) find that follow-on projects receive about 79% as many citations as similar projects that were first to the finding. Scientific innovation is often non-rivalrous because early discoveries generate new opportunities for others in the same domain. Still, in many other settings, one agent's choice can largely diminish the value of that option for others. To capture this, we now introduce rivalry into the model. Specifically, we assume that a project's payoff falls to zero when the number N of agents already selecting that project is greater than $\bar{N}$. This adjustment makes individual payoffs sensitive to competition, with smaller $\bar{N}$ reflecting stronger payoff rivalry. The rest of the model remains unchanged. The results below show how this affects exploration and discovery:

Proposition 4 ("Exploration under Rivalry"). *If payoff rivalry is not extreme (i.e., $\bar{N} > N - A + 1$), then the original weak inequalities still hold under Assumption 1: $\mu|H \leq \mu|M \leq \mu|L \leq \mu|\emptyset$. Moreover, exploration is increasing in rivalry.*

Proof. Without any data, agents still explore all projects in the first period, so $\mu|\emptyset = 1$. If L is revealed ex ante, then agents will explore all the unknown projects in the first period so that $\mu|L = 1$. If H is revealed ex ante, then $\bar{N}$ agents will select the mapped project. The remaining $N - \bar{N}$ agents will randomly select as many as the remaining $A-1$ projects as possible. Therefore, since $N - \bar{N} < A - 1$,

[3]For example, the following parameters satisfy Assumption 1 and Equation (4), and thus lead to the outcome where revealing information about a medium project reduces social welfare and lowers the probability of a breakthrough: $L = 0, M = 6, H = 15, p_L = 7/10, p_M = 1/10, p_H = 2/10, A = 5, N = 5$.

$\mu|H = \frac{N-\bar{N}}{A-1}$. Similarly, if M is revealed ex ante, and Assumption 1 holds, then $\bar{N}$ agents will select the mapped project. The remaining $N - \bar{N}$ agents will randomly select as many as the remaining $A - 1$ projects as possible. Since $N - \bar{N} < A - 1$, then $\mu|M = \frac{N-\bar{N}}{A-1}$. Now, suppose we increase rivalry to $\bar{N} - 1$. Since $N - \bar{N} < A - 1$, when a medium project is revealed, an additional unknown project is explored, and $\mu|M$ increases. In the extreme case, when $\bar{N} = 1$, the revelation of an M project has no impact on exploration as agents will explore all remaining unknown options in the first period. Note this result can be analogously stated in terms of N (instead of $\bar{N}$). Holding $\bar{N}$ constant, if we increase the number of agents to $N + 1$, then if $N - \bar{N} < A - 1$, an additional unknown project is explored, and $\mu|M$ increases. ■

Proposition 5 ("Breakthroughs under Rivalry"). *If payoff rivalry is not extreme (i.e., $\bar{N} > N - A + 1$), then the original strict inequality still holds under Assumption 1: $P(H|M) < P(H|i)$ where $i \in \{\emptyset, L, H\}$. Moreover, breakthrough discoveries are increasing in rivalry.*

Proof. The proof follows directly from the analysis above. If no data is provided, then the probability that H is discovered is still $(1 - (1 - p_H)^A)$. If L is revealed ex ante, then the probability that H is discovered is still $(1 - (1 - p_H)^{(A-1)})$. If H is revealed ex ante, then $P(H|H)$ is still trivially 1. If M is revealed ex ante, the probability that H is discovered is $(1 - (1 - p_H)^{N-\bar{N}})$ and, since $A - 1 > N - \bar{N}$, $P(H|M) < P(H|L) < P(H|\emptyset) < P(H|H)$. Now, suppose we increase rivalry to $\bar{N} - 1$. Since $N - \bar{N} < A - 1$, then an additional unknown project is explored, and $P(H|M)$ increases. Similar to before, this result can also be stated in terms of N. Suppose we increase the number of agents to $N + 1$. Since $N - \bar{N} < A - 1$, then an additional unknown project is explored, and $P(H|M)$ increases. ■

Our key finding is that the streetlight effect persists under modest levels of rivalry but weakens as rivalry increases. Competition pushes agents to explore more, increasing the likelihood of discovering a high-value project. This highlights payoff rivalry as a boundary condition for the streetlight effect.

3 Laboratory Experiment: Design and Results

While our simple theoretical framework helps explain the emergence of the streetlight effect, it remains an open question whether it accurately reflects how agents behave in practice. To explore this, we conducted an online experiment mirroring the structure of our model.

3.1 Experimental Procedure and Logistics

Participants logged into the experimental platform remotely and were assigned to either the data or no-data condition in groups of ten. Upon joining, they received detailed written instructions and watched a compulsory seven-minute video that reiterated the rules and introduced the platform.[4] Participants

[4]The videos shown to participants are available upon request.

were then required to complete a short quiz as an attention and comprehension test. They also had continuous access to the instructions and could contact an experimenter via cell phone or Zoom for support. The experiment consisted of independent "rounds," each following the structure of our theoretical framework. Each round had two periods over which payoffs were calculated. Participants were randomly assigned to groups of five, with groups reshuffled every five rounds. In total, each participant played 20 rounds. At the end of the experiment, we collected demographic information and measured risk preferences using a monetarily incentivized, upscaled version of the Holt and Laury task (Holt and Laury, 2002). Final payments included earnings from one randomly selected round, a show-up fee, and the outcome of the risk elicitation task.

The experiment was programmed using the open-source platform oTree (Chen et al., 2016) and conducted at the Vienna Center for Experimental Economics (VCEE). Participants were recruited from VCEE's subject pool via ORSEE (Greiner, 2015), targeting undergraduate and master's students who had previously participated in no more than five experiments. Participation was voluntary, and individuals could withdraw at any time. We ran 18 sessions with a total of 180 participants, ensuring that no one took part in more than one session. Participants ranged in age from 18 to 52, with an average age of 24.7 years and a standard deviation of 4.7. All sessions were conducted in December 2024. The experimental task lasted approximately 50 minutes, with additional 10 minutes allocated for reading instructions, watching the explanatory video, and completing the attention quiz. Average participant earnings were €15.4, with a standard deviation of €4.6.

3.2 Task Description and Implementation

Participants took on the role of individuals searching for precious gems (Panel A of Figure 1). In each round, they faced five mountains, each hiding one type of gem that could only be revealed through exploration. There were three types of gems, differing in rarity and value: topazes (L), rubies (M), and diamonds (H). While the exact monetary values varied across rounds, diamonds were always more valuable than rubies, and rubies were always more valuable than topazes. Participants were informed that topazes appeared with a 60% probability, rubies with 20%, and diamonds with 20%, though they were not told which gem was hidden behind which mountain. The goal of the game was to find the most valuable gems, as their value directly determined participants' earnings.

In addition to displaying the values and distributions of the gems, the interface tracks the current period and the round number as participants progress through the experiment.[5] Each group of five players

[5]The interface also shows the "block" number, which indicates when participant groups are reshuffled. A new block begins

remains anonymous, and participants cannot interact or communicate directly with one another.[6] Within each round, players take turns selecting a mountain to explore in a randomly determined order that changes every round. A dynamic indicator on the screen highlights when it is their turn to choose. At the start of each round, no player has private information about the locations of the gems, which are randomly reassigned each round (but remain fixed between the two periods of a given round). While waiting for their turn, players can observe which mountains have already been selected. When it is their turn, they are free to choose the same mountain as someone else or a different one.

In the no-data condition, participants begin by selecting one of five mountains to explore in period 1. Once all players have made their choices, the gems hidden in the selected mountains are revealed to everyone, and each player earns the value of the gem from their chosen mountain. In period 2, players again choose from the same mountains, in the same random order, with gem locations unchanged. Now, however, they can see the gems uncovered in period 1 and can either stick with their previous choice or switch to a different mountain. The newly selected mountains are revealed, and their gem values are added to each player's payoff. The data condition follows the same structure, with one key difference: at the start of each round, one mountain is "mapped," and its gem is revealed to all participants. This is the only information available at the outset. Panel B of Figure 1 illustrates this setup. Figure (i) shows the no-data condition, where all mountains are hidden, while Figures (ii), (iii), and (iv) depict the three possible data scenarios, where the revealed mountain contains a low-, medium-, or high-value gem. The revealed mountain is selected by a script using a random sequence.

We collected data from a total of 720 rounds. In 120 of these, participants received data revealing a low-value gem; in 240 rounds, they saw data on a medium-value gem; and in another 120 rounds, the revealed gem was high-value. In the remaining 240 rounds, no initial data about gem locations was provided. We determined the proportion of rounds assigned to each treatment condition based on power calculations. Across the experiment, we used five different combinations of payoff parameters. Specifically, the values for low, medium, and high-value gems were set to one of the following: $(L, M, H) = \{(1, 6, 11), (1, 6, 11.5), (2, 6, 11), (2, 6, 11.5), (3, 7, 12)\}$.

every five rounds, after which players remain in the same group for the next five rounds.

[6] Although participants are aware that their co-players change every five rounds, they are never able to identify who they are playing with. When a player selects a mountain, the others see a message such as "one player selected this mountain," but never learn who made the choice. See Figure 1 for an illustration.

3.3 Results

Group Payoffs. We begin by examining group-level earnings. For each round, we calculate the maximum possible group payoff and express realized group earnings as a percentage of this value. This allows us to compare outcomes across rounds, despite variation in the values and distributions of the low-, medium-, and high-value gems. Panel (i) of Figure 2 plots the average group payoff by condition, comparing the three data treatments to the no-data baseline. Strikingly, revealing data on a medium-value project leads to lower group payoffs than all other conditions, including the case where no data is provided. To quantify these differences, we estimate the following OLS specification:

$$Group\ Payoff_{j,k} = \alpha + \beta Initial\ Data_k + \ \gamma \boldsymbol{X}_k + \epsilon_{j,k}, \tag{5}$$

where $Group\ Payoff_{j,k}$ denotes the payoff for group j in round k, $Initial\ Data_k$ is a categorical variable indicating the type of project revealed at the start of the round, and $\boldsymbol{X}_k$ is a vector of fixed effects that accounts for the session, the specific payoff structure, and the round's position in the session. Standard errors are clustered at the session level. Column 1 of Table 1 presents the results. We find that revealing data on a medium-value mountain reduces group payoffs by approximately 5% relative to the no-data condition, consistent with Proposition 1. Providing data on a high-value mountain increases payoffs by 44.5 percentage points. In contrast, revealing a low-value mountain has no statistically significant effect on group performance.

Group Exploration. Our theoretical framework suggests that partial data on project value can discourage exploration, effectively crowding out data generation. To test this, our next outcome of interest is the share of unmapped mountains explored in a round. Panel (ii) of Figure 2 shows that revealing the location of a medium-value gem significantly reduces exploration. We quantify this using an OLS specification similar to Equation (5), with the dependent variable defined as the share of unmapped mountains explored by the group across both periods.[7] The results in Table 1 show that revealing a high-value gem eliminates the need for exploration, while revealing a low-value gem has no measurable effect. Most notably, revealing a medium-value gem decreases the share of mountains explored by 38.6 percentage points relative to the no-data condition (Column 2). This provides a clear demonstration of the streetlight effect: data can shift the balance from exploration to exploitation, ultimately reducing social welfare by leaving participants stuck on a suboptimal outcome.

Group Breakthroughs. The final outcome of interest is the likelihood that participants discover the high-value option. Panel (iii) of Figure 2 shows that revealing the location of a medium-value gem

[7]Note that there are four unmapped mountains in each of the three data conditions and five in the no-data condition.

significantly lowers the chances of a breakthrough. We quantify this effect using a linear probability model based on the specification in equation (5), with the dependent variable indicating whether a group discovers a high-value gem. Since not all rounds contain a diamond, we limit the analysis to rounds where at least one high-value gem is present. As shown in Column 3 of Table 1, revealing a medium-value mountain reduces the likelihood of discovering the maximum by 56% compared to the no-data condition. In contrast, we find no such reduction when the revealed data points to a low- or high-value gem. Taken together, these results support the predictions of Proposition 3: while data can increase payoffs when it points to the best option, it can also impose substantial societal costs depending on the underlying payoff structure.

The Impact of Competition. Our theory shows that the presence of payoff competition can reduce the intensity of the streetlight effect. We test this experimentally by varying $\bar{N}$, the number of players who can choose a mountain before payoffs fall to zero. The results from the baseline case, presented earlier, implicitly correspond to $\bar{N} = 5$, where players can choose without penalty. We then examine two more conditions: intermediate rivalry ($\bar{N} = 3$) and extreme rivalry ($\bar{N} = 1$). In the intermediate case (Panel A, Table 2), the streetlight effect weakens but does not disappear. Revealing the medium option no longer affects payoffs, but still reduces exploration by roughly 20 percentage points and the likelihood of a breakthrough by 24.5 percentage points (significant at the 10% level). Revealing the high option still increases payoffs and reduces exploration, while revealing a low option continues to have no effect. Under extreme rivalry (Panel B, Table 2), initial data has no significant impact on payoffs, exploration, or breakthroughs. Consistent with Propositions 4 and 5, the streetlight effect declines with increased rivalry and disappears when payoff competition is strongest.

4 Empirical Application: The Genetic Roots of Human Diseases

The preceding sections formalized and tested how the streetlight effect can emerge in lab-based search tasks. We now turn to an empirical application that shows how our framework helps explain real-world patterns in scientific research.

4.1 Setting

Our application examines biomedical research, focusing on scientists' efforts to identify genetic mutations that cause human diseases (see Appendix B for details). Genes carrying causal mutations can serve as drug targets, substantially improving the chances of developing effective treatments (Nelson et al., 2015). However, finding breakthrough targets is a complex search problem: there are over

19,000 protein-coding human genes, each potentially a drug target. In practice, scientists must choose between further investigating known genetic targets or exploring novel candidates. Despite individual incentives to establish priority in new areas (Bobtcheff et al., 2017; Hill and Stein, 2025a), exploration across the genetic space has remained surprisingly limited (Edwards et al., 2011). Research continues to focus on a subset of human genes, a puzzling pattern given widespread recognition that promising drug targets may lie among less-studied genes (Stoeger et al., 2018). One explanation, echoing the streetlight effect, is that earlier data on seemingly promising—but ultimately unproductive—genes have focused scientists' efforts away from exploring more valuable alternatives (Haynes et al., 2018).

To illustrate this, consider two examples of genetic disorders described in Figure 3. As noted in the introduction, research on Tangier disease followed a revealing trajectory. A 1982 study identified a moderate link to the APOA1 gene, which attracted subsequent attention and diverted exploration away from alternative candidates. However, Tangier disease is actually caused by mutations in the ABCA1 gene, which impair the production of functional HDL-C particles. This genetic target was only discovered in 1999. In contrast, the search for the cause of Gardner syndrome, a genetic colon polyposis, unfolded differently. Early investigations yielded only weak associations, prompting a broader search effort. This eventually led to the discovery of mutations in the APC gene, a tumor suppressor that plays a central role in controlling cell growth and is strongly linked to the condition. The APC discovery happened in 1991, eight years before the key breakthrough in Tangier disease, despite both diseases receiving a similar number of publications. These contrasting case studies highlight the streetlight effect in action: the disease that initially showed clearer research progress reached its breakthrough much later.

Building on these cases, we turn to a systematic empirical investigation. Our central proposition is that early discoveries of moderate promise can narrow scientific focus and slow the identification of true genetic drivers. In contrast, weaker early findings tend to promote broader exploration and accelerate discovery. The parallels to our theoretical framework are clear: just as agents in our model search for valuable projects or participants in the lab look for gems hidden in mountains, scientists navigate a vast genetic landscape in pursuit of scientific breakthroughs.

4.2 Data

DisGeNET Database. We compile a dataset of genetics research from 1980 to 2019 using DisGeNET (v7.0), a comprehensive database of gene–disease links drawn from curated sources and PubMed-indexed publications (Piñero et al., 2020; Tranchero, 2025). Because DisGeNET does not include

author information, we supplement it with disambiguated data from Author-ity 2018 (Torvik and Smalheiser, 2021). Additional details on both data sources are provided in Appendix B. Our analysis focuses on articles investigating associations between protein-coding genes and diseases, syndromes, or abnormalities with clear health relevance. For each disease, we record the number of publications along with information on the novel genetic candidates identified each year. To filter out conditions unlikely to have a genetic basis, we restrict the sample to diseases with at least 10 publications over the study period, but results are robust to different cut-offs. The final dataset captures the search and discovery trajectories of 5,519 diseases over a 40-year span.

Measuring the Scientific Value of Genetic Discoveries. Scientists aim to identify genes of high scientific value for each disease. Mirroring our theoretical setup, we classify genetic candidates for a disease into three categories: weak targets (L), middle-value leads (M), and breakthroughs (H). We rely on the score provided by DisGeNET for each gene–disease pair, which ranges from 0 to 1 and summarizes the strength of the available scientific evidence. The score incorporates the number of supporting sources weighted by their credibility, with curated information receiving the greatest weight. We provide extensive details on the DisGeNET score and its features in Appendix B.3. For interpretability, we express a gene–disease pair's scientific value as its percentile within the overall score distribution. Genes below the 60$^{\text{th}}$ percentile are classified as low value, those between the 60$^{\text{th}}$ and 90$^{\text{th}}$ percentiles as medium value, and those above the 90$^{\text{th}}$ percentile as high value. These categories closely align with real-world indicators of therapeutic relevance: clinical citations, approved patents, and granted drugs all increase monotonically with our score categories (Appendix Figure B.1).

Genetic Data Available to Scientists. Our objective is to assess how information on the scientific value of gene candidates shapes subsequent exploration patterns for a given disease. We build on the idea that early discoveries provide data that scientists can choose to exploit through repeated studies, rather than search for new candidates. We define the early search window as the period from 1980 to 2000, which marks the first half of our sample and accounts for just 10% of all publications, during which scientists began identifying potential gene targets. For each disease, we record the highest-scoring gene candidate identified during this period, classifying it as low (L), medium (M), or high (H) based on the categories described above. This captures the state of genetic knowledge available to researchers as of 2000. We then examine how the nature of this early data shapes research activity in the second half of the sample period (2000–2019). Figure C.1 provides a stylized overview of this empirical setup.

Dependent Variables. We construct a dataset at the disease level to examine how cross-sectional differences in early data shape subsequent exploration. Our first dependent variable captures whether scientists identified a gene-disease pair with a high DisGeNET score, corresponding to a breakthrough discovery at the group level in our experimental setup. The second dependent variable measures the number of new gene candidates explored after the early search window, allowing us to assess how the scientific promise of early data constrains the diversity of follow-on research. To account for variation in research intensity across diseases, we normalize this variable by dividing the number of new genes explored by the number of publications. In practice, this captures changes in the average number of new genes explored per disease. The third dependent variable measures the number of years required to reach a breakthrough, defined as the number of years since 1980 (the start of our sample period). This group-level delay offers a concrete indication of the societal cost imposed by the streetlight effect. We include the total number of publications focused on each disease as a control to account for variation in research effort. In addition, DisGeNET assigns each condition a set of disease classes based on the MeSH vocabulary, and our data include 536 unique disease class combinations. Disease class captures features such as whether a condition is congenital or acquired. We include disease-class fixed effects to control for unobserved characteristics shared by similar diseases, and cluster standard errors at the disease-class level to account for correlations across related conditions.

Summary Statistics. Table 3 reports descriptive statistics for the 5,519 diseases in our sample. By the year 2000, 10% of diseases show early data pointing to an L target, 32% to an M target, and the remainder to an H target. The H category is less informative for our purposes, as a breakthrough has already occurred in the early exploration window.[8] On average, it takes 21.8 years to identify a high-value genetic target for a disease by the end of the sample period. Each disease is linked to approximately 295 publications, involving 186 unique principal investigators (PIs), and associated with the discovery of about 130 genes.

5 Empirical Results

5.1 Cross-Sectional Evidence

We begin by examining how the likelihood of a breakthrough varies with early scientific data, comparing outcomes for diseases with information only on low-value genes to those with data on genes of medium or high value. To do this, we estimate the following cross-sectional OLS specification at the

[8]Note that we do not include a "no data" condition here, as most diseases had seen some level of investment before 2000.

disease level:

$$Breakthrough\ (0/1)_i = \alpha + \beta(Max\ Found:\ X_i) + \ \gamma \boldsymbol{X}_i + \epsilon_i, \tag{6}$$

where $Breakthrough\ (0/1)_i$ equals 1 if at least one publication discovers a genetic target with a high DisGeNET score for disease i, and 0 otherwise. The variable $(Max\ Found:\ X_i)$ is a categorical indicator for the highest DisGeNET score identified in the early search window, classified as L, M, or H. $\boldsymbol{X}_{i,t}$ is a vector of controls that includes the number of publications on the disease, as a proxy for search efforts, and fixed effects for disease class, taking into account broader genetic similarities between related diseases. The results are reported in Panel A of Table 4. While early data on a high-value genetic target mechanically increases the likelihood of a breakthrough, the more interesting comparison lies with medium-value targets. Diseases with early data on medium-value genes are 11 percentage points less likely to experience a breakthrough than those with only low-value initial findings.

One possible explanation for this counterintuitive finding is that the early discovery of a promising—but ultimately suboptimal—genetic target diverts attention from the search for a true breakthrough. Column 2 of Table 4 presents evidence consistent with this mechanism. Using the same specification as in Equation (6), we find that early data on a medium-value target reduces exploration of new genes by almost 20 percentage points. Notably, this drop is nearly half as large as the effect of an early breakthrough itself. Column 3 quantifies the real-world cost of reduced exploration. Identifying a medium-value target early on delays the eventual breakthrough by 1.7 years, which corresponds to an increase of 8% relative to the sample mean of 21.7 years.

Our theoretical framework and behavioral experiments suggest that the streetlight effect should weaken as rivalry increases. Does this prediction hold in our empirical setting? In our experiments, we can directly manipulate the threshold $\bar{N}$, which represents the number of individuals who can benefit from a project before the payoff erodes. Here, we take $\bar{N}$ as fixed, assuming it is broadly similar across diseases due to common scientific norms of credit allocation. Still, our comparative statics in Propositions 4 and 5 show that even with a constant $\bar{N}$, increasing N (i.e., the number of competitors) should reduce the streetlight effect. This insight allows us to proxy rivalry empirically using the number of scientists who have studied each disease, echoing recent work on competition in science (Hill and Stein, 2025a). We classify diseases as more or less competitive based on the number of active PIs. Using our cross-sectional specification from Equation (6), we run split-sample regressions for diseases in the top and bottom quartiles of this distribution.

The results of this analysis are presented in Panel B of Table 4. In Column 1, where we restrict the sample to diseases with fewer scientists, we find that early data on a medium-value target reduces the likelihood of a breakthrough by 17 percentage points. By contrast, for diseases with more competition (Column 2), there is no significant change in breakthrough likelihood. A similar pattern holds for exploration activity: early data reduces exploration of new genes by 20% relative to the sample mean in diseases with fewer scientists (Column 3), but this effect disappears when more PIs are engaged in the search for genetic roots (Column 4). Finally, early data on a tempting genetic target increases the time to breakthrough in diseases with fewer researchers involved (Column 5), while the effect is not significant in more competitive settings (Column 6). Taken together, these results suggest that competition helps offset the streetlight effect that uneven data availability might create.

5.2 Instrumental Variables (IV)

The empirical patterns are consistent with our theorization of the streetlight effect and echo concerns raised by scientists about the lack of exploration in this field (Haynes et al., 2018; Stoeger et al., 2018). However, issues about causality remain, since the generation of early data reflects scientists' endogenous exploration choices. As a first step, we note that including fixed effects for disease classes helps control for unobserved characteristics shared across related diseases. Yet, certain disease-specific features could still correlate with the nature and volume of early scientific data, potentially driving our results.

To help rule out this concern and bolster the causal interpretation of our findings, we leverage the fact that many human genes have orthologous counterparts—that is, genes in other species that share a common ancestor gene and thus retain similar biological sequences and functions. Scientists frequently use animals as models to experimentally study human orthologs at lower cost and with fewer ethical constraints (Li et al., 2017). In particular, genes with orthologs in the commonly used laboratory mice tend to receive more attention from scientists out of sheer convenience (Stoeger et al., 2018). We retrieve information on gene orthology from the National Center for Biotechnology Information (NCBI). In our data, human genes with mouse orthologs appear 2.6 years earlier in scientific publications and are about 27% more likely to have been explored before the year 2000 (Figure 4, Panel A). This confirms that researchers often prioritize these genes, and within a disease, their discoveries emerge earlier (Appendix Table C.1). Yet, nothing ensures that orthologs are equally relevant for every disease. Since the strength of any association between these overlapping genes and a given disease is effectively exogenous, delays stemming from researchers focusing on medium-value

orthologous genes can be more credibly attributed to convenience rather than to unobserved disease characteristics.

Building on this intuition, we construct an instrumental variable based on the distribution of orthologous gene candidates. For each disease, we measure the share of orthologous genes classified as medium-value (M) candidates (see Appendix Figure C.2 for a stylized visualization). If the orthologous gene pool contains more medium-value targets for a particular disease, scientists should be more likely to encounter a medium-value discovery early in their exploration. Indeed, our instrument exogenously shifts the probability of identifying a medium-value gene, as confirmed by a strong first-stage regression (Figure 4, Panel B). For example, Tangier Disease has 27 gene candidates with mouse orthologs, 7 of which are medium-value ($M\ share_{Tangier}$ = 26%); APA01, an M ortholog, was indeed discovered early, in 1982. In contrast, only 2 of Gardner Syndrome's 23 orthologous genes are of medium-value ($M\ share_{Gardner}$ = 8%), leading scientists to identify the causal APC gene without distraction by medium-value discoveries. Figure C.3 supports this logic across our broader sample of diseases, showing reduced-form evidence linking the M share of ortholog genes to breakthrough likelihood, exploration extent, and discovery delay.

The results of our IV analysis are presented in Table 5. We replicate the same three cross-sectional specifications as before, but now instrument for $(MaxFound : M)$ using the share of ortholog genes corresponding to an M target. In this setup, the 2SLS coefficients can be interpreted as a local average treatment effect (LATE), capturing the effect in the subset of diseases for which the instrument shifts the likelihood of identifying an M target. Column 1 presents the first stage of our IV, showing that the M share of ortholog genes is strongly associated with the highest-scoring gene association being M, with an F-statistic of 154. Columns 2 through 4 report the second-stage results for our three main outcomes. Consistent with our earlier findings, we find that early data on a medium-value target reduces the likelihood of a breakthrough (Column 2), decreases the number of new genes explored (Column 3), and increases the delay to a breakthrough (Column 4). This analysis confirms that the patterns observed in genetics stem from the streetlight effect created by early data (Haynes et al., 2018).

5.3 Exploration Dynamics

Next, we offer additional evidence to bolster confidence in the mechanism proposed by our theory. If early data on medium-value genetic targets indeed crowds out exploration, we should see a drop in research efforts aimed at discovering new genes in the years following a medium-value gene. To test this idea, we construct a panel at the disease-year level. While our earlier analyses relied on cross-

sectional estimates at the disease level, this alternative approach allows us to track how exploration patterns change over time. For each disease, we count the number of new genes investigated in a given year, along with the total number of publications as a proxy for research effort. Appendix Table C.2 presents descriptive statistics at the disease-year level. On average, each disease receives 7.4 publications per year focused on its genetic underpinnings, typically leading to the exploration of 3.3 new genes.

We then estimate the following event study specification using OLS:

$$Group\ Exploration_{i,t} = \alpha + \sum_{z} \beta_t Medium\ Gene_i \times 1(z)\ +\ \gamma \boldsymbol{X}_{i,t} + \epsilon_{i,t}, \tag{7}$$

where $Group\ Exploration_{i,t}$ denotes the number of new genes explored for disease i in year t, normalized by the number of articles published. $Medium\ Gene_i \times 1(z)$ the number of years that have elapsed since a medium-value association was first discovered for disease i, and $\boldsymbol{X}_{i,t}$ is a vector of controls that include disease fixed effects, year fixed effects, and the number of papers published each year. For the small number of diseases with multiple medium-value genes, we define the time lags relative to the discovery of the first one. To account for the mechanical uptick in exploration during the year of discovery, we exclude the focal gene and its corresponding publication from our calculations, but our results are robust to their inclusion.

Panel A of Figure 5 plots the regression coefficients. The results show an immediate, significant, and persistent drop in exploration following the discovery of a medium-value genetic target. Reassuringly, there is no evidence of pre-trends, suggesting that the observed decline is indeed driven by the discovery itself. In Panel B, we re-estimate the specification in Equation (7) and find a similar pattern: research efforts on new genes also decline after the discovery of a high-value target. Appendix Table C.5 reports the corresponding estimates from a difference-in-differences specification. We find that yearly exploration of new genes drops by 24% relative to the sample mean after a medium-value target is identified. The effect is even larger following the discovery of a high-value target, with exploration falling by around 35%. Taken together with the IV results, these estimates offer additional support for the predictions of our theoretical framework.

5.4 Robustness

We assess the robustness of our results by relaxing several key choices in the main specification. First, while we excluded diseases with very few publications to focus on those more likely to have genetic roots, our results hold under alternative sample cut-offs (Appendix Table C.6). Similarly, excluding

the top 1% of most-studied diseases does not change the findings (Appendix Table C.7). Second, we defined payoffs based on percentiles of the DisGeNET score. While our definitions map into real-world outcomes (Appendix Figure B.1), changing the percentiles used to define an M genetic discovery does not affect our results (Appendix Tables C.8 and C.9). Third, redefining the early search window yields consistent results (Appendix Table C.10). Appendix Table C.11 shows robustness to a disease-specific definition, where "early" refers to the years before the first 10% of publications for each disease. Finally, the findings remain stable under alternative windows for tracking exploration dynamics following an M discovery (Appendix Table C.12).

One potential concern is that focusing on a medium-value gene could be a rational choice when there is ambiguity about whether a high-value target exists at all. This might partly explain the drop in exploration following the discovery of an M. To address this, we draw on the genetic relationships between diseases. The MeSH vocabulary defines hierarchical linkages between diseases based on shared etiology, biological mechanisms, and other biomedical features. Using this classification, we restrict the analysis to diseases closely related to conditions where a breakthrough (H) has already occurred. In these cases, the existence of valuable targets is less ambiguous, as related diseases often share underlying biological processes.[9] Re-estimating the event study specification in Equation (7), we find consistent results: as shown in Figure C.4, data on a medium-value gene still dampens exploration, even within this subset of diseases.

Relatedly, it is possible that our cross-sectional results reflect the absence of valuable genetic targets, rather than suboptimal exploration behavior. To test this, we narrow our analysis to diseases that had a high-value genetic association identified by 2019. These results are presented in Appendix Table C.13. While this restriction prevents us from estimating effects on group breakthroughs, we still observe longer delays for diseases where early data pointed to an M-value target. We find no change in exploration over the full sample period, likely because all diseases in this sample eventually saw a breakthrough and, by definition, received some level of exploration. Still, we detect a significant decline in exploration activity in the years immediately following the early M discovery.

Finally, one reason scientists might continue to focus on M candidates is the prospect of positive spillovers that could benefit research on related diseases. These spillovers could come from comple-

[9]For instance, Ulcerative Colitis [MeSH tree code: C06.405.469.432.249] and Crohn's Disease [MeSH tree code: C06.405.469.432.500] are "sibling" sub-branches of the "parent" disease Inflammatory Bowel Diseases [MeSH tree code: C06.405.469.432]. Once a gene is identified as a high-value target for Crohn's there is a higher chance that a breakthrough exists for Ulcerative Colitis (which could either be the same gene or another one).

mentary insights, such as new methods or a deeper understanding of protein function and genetics. If genes classified as M in one disease often end up as H candidates in related diseases, then continued focus on them might be rational. To evaluate this possibility, we test whether a gene is more likely to be a breakthrough for a given disease when it is classified as an M in a related disease. The results are presented in Panel A of Appendix Table C.14. The effect is small and only slightly larger than when the gene is classified as L, and much smaller than when it is classified as H in a sibling disease. By contrast, as Panel B shows, genes classified as M for one disease are likely to remain M in related diseases. This suggests that spillovers are limited in scope and are unlikely to justify continued attention to M candidates.

6 Conclusion

In this paper, we examine the paradoxical role of data provision in shaping innovative search, a dynamic we refer to as the "streetlight effect." Our theoretical model shows that access to partial data on past successes can narrow the search space and trigger free-riding, ultimately reducing the diversity of exploration and hampering breakthrough discoveries. This prediction is supported by our empirical findings. In our lab experiments, revealing data on a medium-value project lowered group payoffs by 5% and reduced the likelihood of a breakthrough by 56% compared to the no-data condition. We extend this analysis using observational data from scientific research on the genes responsible for human diseases. Our approach includes multiple research designs, including an instrumental variable strategy based on exogenous genetic overlaps between human and mouse genes. The results show that diseases with early data on a middle-value target are, on average, 16 percentage points less likely to yield breakthroughs, with discoveries delayed by nearly three years due to reduced exploration. We also find that payoff competition moderates these effects by lowering the attractiveness of known options and breaking the cycle of low data generation. Taken together, our theoretical, experimental, and empirical evidence highlights how the streetlight effect shapes the direction of innovative search.

Our findings challenge the conventional belief that more data is always better for innovation. When data is incomplete and narrowly focused, as in our setting, it can unintentionally steer researchers toward suboptimal projects. Our evidence from genetics highlights how this pattern can emerge endogenously in decentralized and parallel exploration endeavors such as scientific research. This has important implications for policymakers and funding agencies involved in data creation and dissemination, whose goal should be to provide broad "floodlights" that illuminate the entire search space. Our findings reinforce the value of publicly funded, comprehensive mapping initiatives such as the Human Genome

Project (Williams, 2013) and Landsat satellite imagery (Nagaraj, 2022), which serve as shared data infrastructure for scientific discovery. They also highlight the importance of strengthening institutions such as the U.S. Census Bureau's FSRDCs (Nagaraj and Tranchero, 2024), which enable research access to existing large-scale datasets at relatively low public cost.

For individual innovators, the key takeaway is that past data should be treated as a strategic input rather than followed blindly. In environments where data is uneven or incomplete, setting aside existing information can promote breakthrough innovation. Our findings lend support to corporate practices like skunkworks, where firms intentionally restrict the internal diffusion of early R&D results. They also underscore the value of delaying the release of intermediate project information unless there is strong evidence that the project represents a high-value lead (Boudreau and Lakhani, 2015). More broadly, as innovation and decision-making become increasingly data-driven, it is important to recognize that technologies like AI are often trained on uneven historical data. This can inadvertently narrow the scope of exploration by reproducing the streetlight effect (Kim, 2023). While existing work has focused on the risk of false positives in AI predictions (Tranchero, 2024), our evidence suggests that the risk of false negatives in data-driven innovation may be even greater. At the same time, AI enables initiatives like AlphaFold, which provide broad and unfiltered predictions supporting discovery beyond the bounds of known data. Understanding the nuanced implications of AI for innovation is an exciting direction for future research.

While our study draws strength from combining theoretical modeling, laboratory experimentation, and empirical analysis, there remain several opportunities for further improvement. One direction would be to extend the current two-period framework into a continuous learning model, which would better capture the iterative and dynamic nature of innovation. Our model could also be extended to explore how control rights in organizations might help coordinate search efforts and prevent herding (Aghion et al., 2008; Arora et al., 2025). Another promising avenue lies in broadening our definition of data to include dimensions such as precision, informativeness, and bias, all of which are likely to shape search behavior in meaningful ways. The observational analysis, while strengthened by an instrumental variable approach, could also be complemented by research designs that introduce direct experimental variation in the data provided. Expanding the analysis to consider a broader set of innovation outcomes across diverse domains would further enhance the generalizability of our findings.

References

Aghion, P., M. Dewatripont, and J. C. Stein (2008): "Academic freedom, private-sector focus, and the process of innovation," *The RAND Journal of Economics*, 39, 617–635.

Agrawal, A., J. McHale, and A. Oettl (2024): "Artificial intelligence and scientific discovery: A model of prioritized search," *Research Policy*, 53, 104989.

Arora, A., S. Hasan, and W. D. Miles (2025): "If you had one shot: Scale and herding in innovation experiments," *NBER Working Paper 33682*.

Azoulay, P., J. S. Graff Zivin, and G. Manso (2011): "Incentives and creativity: Evidence from the academic life sciences," *The RAND Journal of Economics*, 42, 527–554.

Bergemann, D. and A. Bonatti (2019): "Markets for information: An introduction," *Annual Review of Economics*, 11, 85–107.

Bergemann, D. and J. Valimaki (2008): "Bandit problems," in *The New Palgrave Dictionary of Economics, 2nd ed.*, Macmillan Press, 336–340.

Bessen, J., S. M. Impink, L. Reichensperger, and R. Seamans (2022): "The role of data for AI startup growth," *Research Policy*, 51, 104513.

Bloom, N., C. I. Jones, J. Van Reenen, and M. Webb (2020): "Are ideas getting harder to find?" *American Economic Review*, 110, 1104–1144.

Bobtcheff, C., J. Bolte, and T. Mariotti (2017): "Researcher's dilemma," *The Review of Economic Studies*, 84, 969–1014.

Bolton, P. and C. Harris (1999): "Strategic experimentation," *Econometrica*, 67, 349–374.

Bordalo, P., N. Gennaioli, Y. Ma, and A. Shleifer (2020): "Overreaction in macroeconomic expectations," *American Economic Review*, 110, 2748–2782.

Bordalo, P., N. Gennaioli, and A. Shleifer (2012): "Salience theory of choice under risk," *The Quarterly Journal of Economics*, 127, 1243–1285.

——— (2013): "Salience and consumer choice," *Journal of Political Economy*, 121, 803–843.

Boudreau, K. J. and K. R. Lakhani (2015): "'Open' disclosure of innovations, incentives and follow-on reuse: Theory on processes of cumulative innovation and a field experiment in computational biology," *Research Policy*, 44, 4–19.

Boyce, J. R., D. M. Bruner, and M. McKee (2016): "Strategic experimentation in the lab," *Managerial and Decision Economics*, 37, 375–391.

Cao, R., R. Koning, and R. Nanda (2024): "Sampling bias in entrepreneurial experiments," *Management Science*, 70, 7283–7307.

Chen, D. L., M. Schonger, and C. Wickens (2016): "oTree—An open-source platform for laboratory, online, and field experiments," *Journal of Behavioral and Experimental Finance*, 9, 88–97.

Cockburn, I. M., R. Henderson, and S. Stern (2019): "The impact of artificial intelligence on innovation," *The Economics of Artificial Intelligence: An Agenda*, 115–152.

Ederer, F. and G. Manso (2013): "Is pay for performance detrimental to innovation?" *Management Science*, 59, 1496–1513.

Edwards, A. M., R. Isserlin, G. D. Bader, S. V. Frye, T. M. Willson, and F. H. Yu (2011): "Too many roads not taken," *Nature*, 470, 163–165.

Ewens, M., R. Nanda, and M. Rhodes-Kropf (2018): "Cost of experimentation and the evolution of venture capital," *Journal of Financial Economics*, 128, 422–442.

Farboodi, M. and L. Veldkamp (2020): "Long-run growth of financial data technology," *American Economic Review*, 110, 2485–2523.

Gates, A. J., D. M. Gysi, M. Kellis, and A.-L. Barabási (2021): "A wealth of discovery built on the Human Genome Project—by the numbers," *Nature*, 590, 212–215.

Gordon, R. (2016): *The rise and fall of American growth: The US standard of living since the civil war*, Princeton University Press.

Greiner, B. (2015): "Subject pool recruitment procedures: organizing experiments with ORSEE," *Journal of the Economic Science Association*, 1, 114–125.

Haynes, W. A., A. Tomczak, and P. Khatri (2018): "Gene annotation bias impedes biomedical research," *Scientific Reports*, 8, 1362.

Henrich, J., S. J. Heine, and A. Norenzayan (2010): "Most people are not WEIRD," *Nature*, 466, 29–29.

Henry, E., M. Loseto, and M. Ottaviani (2022): "Regulation with experimentation: Ex ante approval, ex post withdrawal, and liability," *Management Science*, 68, 5330–5347.

Hill, R. and C. Stein (2025a): "Race to the bottom: Competition and quality in science," *The Quarterly Journal of Economics*, 140, 1111–1185.

——— (2025b): "Scooped! Estimating rewards for priority in science," *Journal of Political Economy*, 133.

Hoelzemann, J. and N. Klein (2021): "Bandits in the lab," *Quantitative Economics*, 12, 1021–1051.

——— (2025): "Breakdowns in the lab," *University of Vienna and University of Montreal*.

Holt, C. A. and S. K. Laury (2002): "Risk aversion and incentive effects," *American Economic Review*, 92, 1644–1655.

Hörner, J., N. Klein, and S. Rady (2022): "Overcoming free-riding in bandit games," *The Review of Economic Studies*, 89, 1948–1992.

Jones, C. I. and C. Tonetti (2020): "Non-rivalry and the economics of data," *American Economic Review*, 110, 2819–2858.

Kao, J. (2024): "Charted territory: Mapping the cancer genome and R&D decisions in the pharmaceutical industry," *UCLA Anderson*.

Kehoe, A. and V. Torvik (2019): "Predicting controlled vocabulary based on text and citations: Case studies in medical subject headings in MEDLINE and patents," *University of Illinois at Urbana-Champaign*.

Keller, G., S. Rady, and M. Cripps (2005): "Strategic experimentation with exponential bandits," *Econometrica*, 73, 39–68.

Kerr, W. R., R. Nanda, and M. Rhodes-Kropf (2014): "Entrepreneurship as experimentation," *Journal of Economic Perspectives*, 28, 25–48.

Kim, S. (2023): "Shortcuts to innovation: The use of analogies in knowledge production," *Columbia Business School*.

Klein, N. and S. Rady (2011): "Negatively correlated bandits," *The Review of Economic Studies*, 78, 693–732.

Krieger, J. L. (2021): "Trials and terminations: Learning from competitors' R&D failures," *Management Science*, 67, 5525–5548.

Levinthal, D. A. (1997): "Adaptation on rugged landscapes," *Management Science*, 43, 934–950.

Li, D., P. Azoulay, and B. N. Sampat (2017): "The applied value of public investments in biomedical research," *Science*, 356, 78–81.

Manso, G. (2011): "Motivating innovation," *The Journal of Finance*, 66, 1823–1860.

——— (2016): "Experimentation and the returns to entrepreneurship," *The Review of Financial Studies*, 29, 2319–2340.

March, J. G. (1991): "Exploration and exploitation in organizational learning," *Organization Science*, 2, 71–87.

Nagaraj, A. (2022): "The private impact of public data: Landsat satellite maps increased gold discoveries and encouraged entry," *Management Science*, 68, 564–582.

Nagaraj, A. and M. Tranchero (2024): "How does data access shape science? The impact of federal statistical research data centers on economics research," *NBER Working Paper 31372*.

Nelson, M. R., H. Tipney, J. L. Painter, J. Shen, P. Nicoletti, Y. Shen, A. Floratos, P. C. Sham, M. J. Li, J. Wang, et al. (2015): "The support of human genetic evidence for approved drug indications," *Nature Genetics*, 47, 856–860.

Nguyen, D.-T., S. Mathias, C. Bologa, S. Brunak, N. Fernandez, A. Gaulton, A. Hersey, J. Holmes, L. J. Jensen, A. Karlsson, et al. (2017): "Pharos: Collating protein information to shed light on the druggable genome," *Nucleic Acids Research*, 45, D995–D1002.

Olea, J. L. M. and C. Pflueger (2013): "A robust test for weak instruments," *Journal of Business & Economic Statistics*, 31, 358–369.

Piñero, J., J. M. Ramírez-Anguita, J. Saüch-Pitarch, Ronzano, et al. (2020): "The DisGeNET knowledge platform for disease genomics: 2019 update," *Nucleic Acids Research*, 48, D845–D855.

Stoeger, T., M. Gerlach, R. I. Morimoto, and L. A. Nunes Amaral (2018): "Large-scale investigation of the reasons why potentially important genes are ignored," *PLoS Biology*, 16, e2006643.

Torvik, V. and N. Smalheiser (2021): "Authority 2018—PubMed author name disambiguated dataset," *University of Illinois Urbana-Champaign*.

Tranchero, M. (2024): "Finding diamonds in the rough: Data-driven opportunities and pharmaceutical innovation," *The Wharton School*.

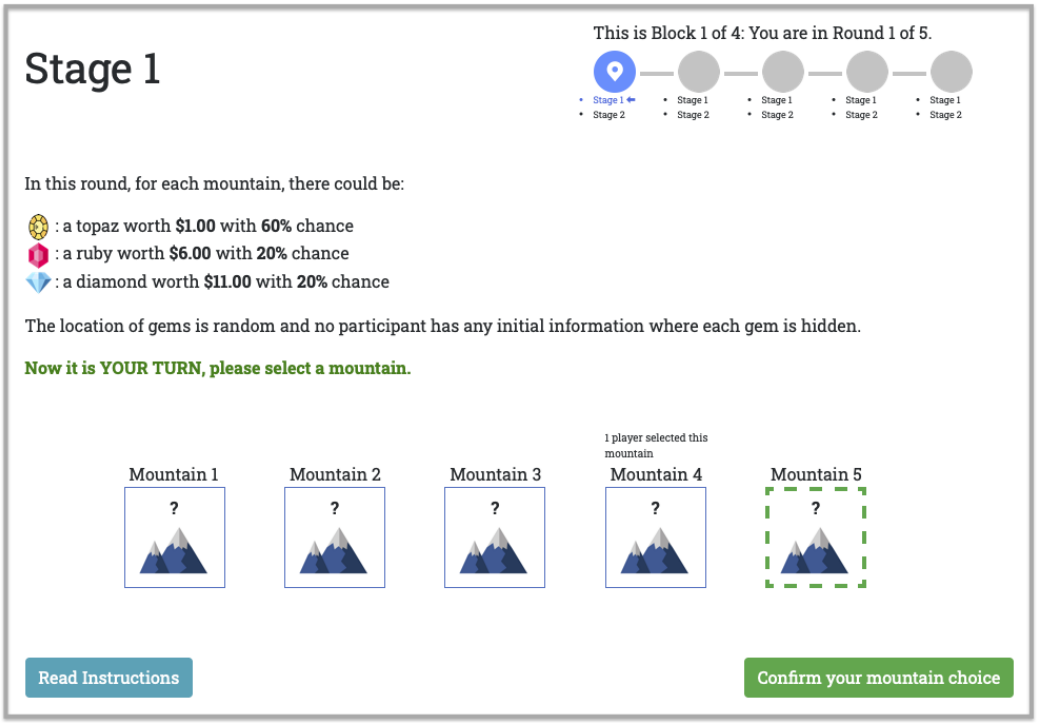
——— (2025): "Data-driven search and the birth of theory: Evidence from genome-wide association studies," *The Wharton School*.

Vermorel, J. and M. Mohri (2005): "Multi-armed bandit algorithms and empirical evaluation," in *Machine Learning: ECML 2005*, ed. by J. Gama, R. Camacho, P. B. Brazdil, A. M. Jorge, and L. Torgo, Berlin, Heidelberg: Springer Berlin Heidelberg, 437–448.

Williams, H. L. (2013): "Intellectual property rights and innovation: Evidence from the human genome," *Journal of Political Economy*, 121, 1–27.

7 Tables and Figures

Panel A: User Interface



Panel B: Examples of No-Data Condition and Data Conditions

(i) No-data condition

(ii) Low-value condition

(iii) Medium-value condition

(iv) High-value condition

Figure 1: Experimental Platform.

Note: This figure shows the interface participants saw during our online experiment. Panel A illustrates the platform as it appeared in the no-data condition. In this example, Mountain 4 was selected by one other participant, while the user chose Mountain 5. Note that the dollar value of the gems changes in every round and is displayed on the left. Panel B presents the four experimental conditions. In the data condition, participants are shown the value of the gem hidden behind one randomly selected mountain—this could be the medium, the high, or one of the low outcomes. .

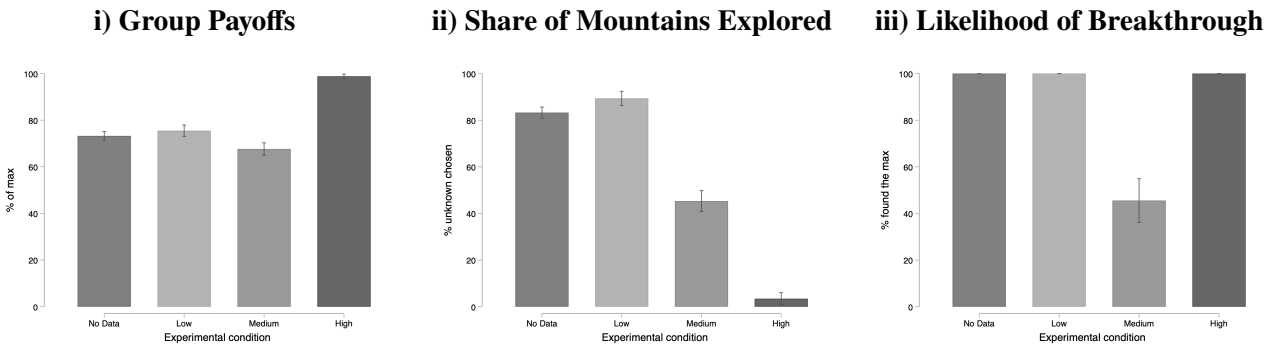


Figure 2: Round Outcomes by Experimental Condition.

Note: Figure (i) displays the average group payoffs per round, by experimental condition. Payoffs are calculated as a share of the maximum possible payoff possible in each round. Figure (ii) shows the average share of unmapped mountains selected per round, by experimental condition. Figure (iii) reports the proportion of rounds in which the maximum payoff was uncovered, by experimental condition. Error bars indicate 95% confidence intervals. See text for more details.

Table 1: Round-Level Experimental Outcomes.

	Group Payoff	Group Exploration	Group Breakthrough
	(1) Group Earnings ($)	(2) Options Explored (%)	(3) Found Maximum (0/1)
High	44.520***	-81.330***	-1.500
	(0.894)	(3.049)	(4.277)
Low	1.545	4.598	-2.045
	(1.228)	(3.120)	(3.631)
Medium	-3.133***	-38.634***	-56.338***
	(0.670)	(2.577)	(5.057)
Session FE	Yes	Yes	Yes
Block order FE	Yes	Yes	Yes
Payoff structure FE	Yes	Yes	Yes
Observations	480	480	364

Note: † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the session level in parentheses.

Estimates from OLS models. The unit of analysis is the group-round level (480 rounds in total). Column 2 includes only the rounds in which at least one diamond was present (364 rounds). In all models, payoffs are non-rival if multiple agents choose the same project. *Group Earnings*= sum of payoffs in a group-round; *Options Explored*= share of unknown mountains explored in the round; *Found Maximum*:0/1=1 if the location of the maximum was found by any participant. The excluded category is the control condition without data. See text for more details.

Table 2: Round-Level Outcomes of the Experiment with Payoff Rivalry.

Panel A: Intermediate Payoff Rivalry

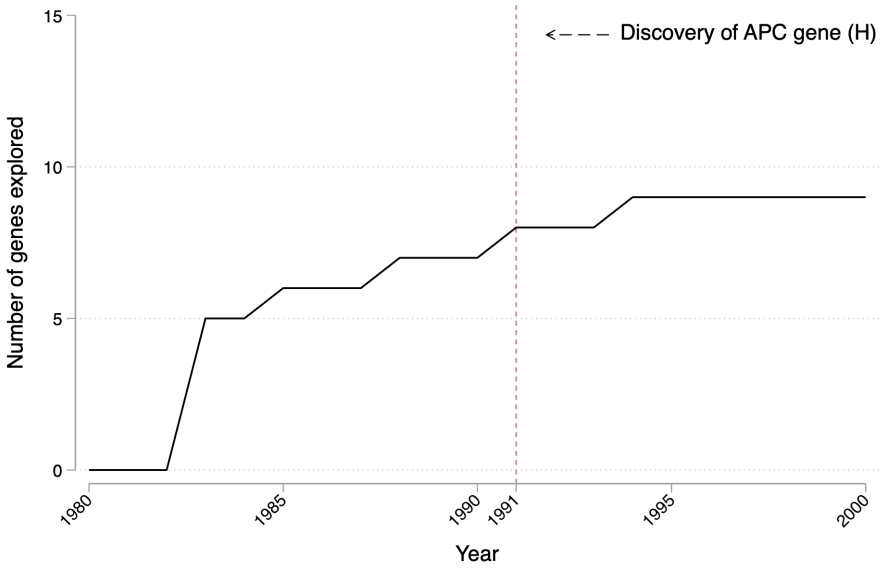
	Group Payoff	Group Exploration	Group Breakthrough
	(1) Group Earnings ($)	(2) Options Explored (%)	(3) Found Maximum (0/1)
High	19.048*	-15.520*	-3.681
	(2.253)	(1.918)	(1.712)
Low	-7.212	8.759†	-4.162
	(2.803)	(2.939)	(5.394)
Medium	-1.118	-19.870***	-24.470†
	(1.714)	(0.524)	(5.726)
Session FE	Yes	Yes	Yes
Block order FE	Yes	Yes	Yes
Payoff structure FE	Yes	Yes	Yes
Observations	120	120	90

Panel B: Extreme Payoff Rivalry

	Group Payoff	Group Exploration	Group Breakthrough
	(1) Group Earnings ($)	(2) Options Explored (%)	(3) Found Maximum (0/1)
High	3.698	0.000	-0.000
	(2.400)	(0.000)	(0.000)
Low	-6.888†	0.000	-0.000
	(1.753)	(0.000)	(0.000)
Medium	3.345*	0.000	-0.000
	(0.364)	(0.000)	(0.000)
Session FE	Yes	Yes	Yes
Block order FE	Yes	Yes	Yes
Payoff structure FE	Yes	Yes	Yes
Observations	120	120	90

Note: † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
Estimates from OLS models. The unit of analysis is the group-round level. In Panel A, payoffs exhibit intermediate rivalry ($\bar{N} = 3$), meaning that once three agents have selected the same mountain in a given period, any additional agents choosing that mountain will receive a payoff of zero. In Panel B, payoffs exhibit extreme rivalry ($\bar{N} = 1$), where only the first agent selecting a mountain earns a positive payoff, while any subsequent agents choosing the same mountain receive a payoff of zero. *Group Earnings*= sum of payoffs in a group-round; *Options Explored*= share of unknown mountains explored in the round; *Found Maximum*:0/1=1 if the location of the maximum was found by any participant. The excluded category is the control condition without data. See text for more details.

Panel A: Gardner's Syndrome



Panel B: Tangier's Disease

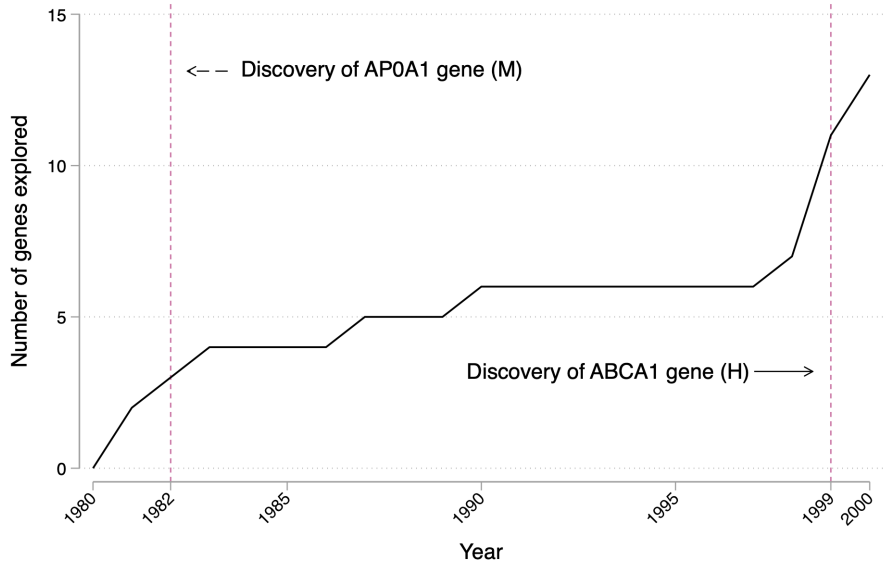


Figure 3: Two Case Studies in Search for the Genetic Origins of Human Diseases.

Note: The solid black line shows the cumulative number of gene candidates explored for the disease up to each year. Panel A displays data for Gardner's syndrome, with the vertical line marking the year the association with the APC gene was discovered (DisGeNET score in the 99th percentile). Panel B shows data for Tangier's disease. The first vertical line marks the discovery of the APA01 association (DisGeNET score in the 60th percentile), while the second marks the discovery of the ABCA1 association (DisGeNET score in the 99th percentile). All other genes explored were below the 60th percentile of the DisGeNET score.

Table 3: Descriptive Statistics of the DisGeNET Database.

	Mean	Median	Sd	Min	Max	N
Max Found: Low (0/1)	0.10	0.00	0.30	0	1	5519
Max Found: Medium (0/1)	0.32	0.00	0.47	0	1	5519
Max Found: High (0/1)	0.58	1.00	0.49	0	1	5519
Year of First Low Score	1991.45	1992.00	5.58	1980	2000	1530
Year of First Medium Score	1993.71	1994.00	7.42	1980	2019	2890
Year of First High Score	1994.98	1995.00	8.13	1980	2019	3964
Delay (Years since 1980)	21.75	18.00	12.82	0	39	5519
Total Publications	294.63	48.00	1983.81	9	94470	5519
Total Genes Discovered	129.95	32.00	394.35	1	8545	5519
New Genes per Paper	0.73	0.72	0.50	0	9	5519
Total PIs on disease	186.34	38.00	1042.30	5	43749	5519

Note: This table presents cross-sectional descriptive statistics for our sample at the disease level. *Max Found: Low*: 0/1=1 if the gene with the highest DisGeNET score found during the early exploration period is classified as L. *Max Found: Medium*: 0/1=1 if the gene with the highest DisGeNET score found during the early exploration period is classified as M. *Max Found: High*: 0/1=1 if the gene with the highest DisGeNET score found during the early exploration period is classified as H. *Year of First Low Score* = the year of the first discovery involving a gene in the L category. *Year of First Medium Score*: the year of the first discovery involving a gene in the M category. *Year of First High Score*: the year of the first discovery involving a gene in the H category. *Delay (Years since 1980)* = the number of years elapsed before any H gene is discovered for the disease. *Total Publications* = the number of publications about the disease during the sample period (1980-2019). *Total Genes Discovered* = the number of genes explored for the disease during the sample period (1980-2019). *New Genes per Publication* = the number of new genes explored per scientific publication during the sample period (1980-2019). *Total PIs* = the number of unique principal investigators (PIs) that have studied the disease during the sample period (1980-2019). See text for more details.

Table 4: Disease-Level Outcomes of Genetic Search.

Panel A: Full Sample

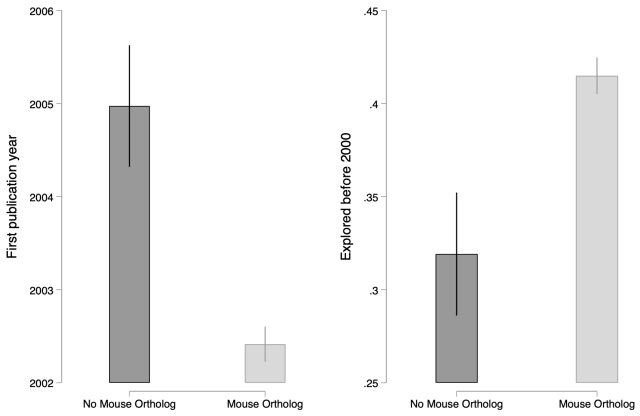
	Group Breakthrough	Group Exploration	Group Delay
	(1) High-Value Gene (0/1)	(2) New Genes/Papers	(3) Years From 1980
Max Found: M	-0.105**	-0.144***	1.743***
	(0.033)	(0.023)	(0.519)
Max Found: H	0.514***	-0.261***	-20.371***
	(0.042)	(0.028)	(0.692)
Disease Class FE	Yes	Yes	Yes
Count of Publications	Yes	Yes	Yes
N	4760	4760	4760

Panel B: Split Samples

	Group Breakthrough		Group Exploration		Group Delay	
	High-Value Gene (0/1)		New Genes/Papers		Years From 1980	
High Competition:	(1) No	(2) Yes	(3) No	(4) Yes	(5) No	(6) Yes
Max Found: M	-0.165***	-0.0439	-0.144**	0.0779	2.492***	2.296
	(0.0433)	(0.154)	(0.0510)	(0.0972)	(0.637)	(2.973)
Max Found: H	0.516***	0.541***	-0.132†	-0.00319	-18.66***	-21.45***
	(0.0470)	(0.159)	(0.0790)	(0.0949)	(0.674)	(2.992)
Disease Class FE	Yes	Yes	Yes	Yes	Yes	Yes
Count of Publications	Yes	Yes	Yes	Yes	Yes	Yes
N	1106	1236	1106	1236	1106	1236

Note: † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.
Standard errors clustered at the disease-class level in parentheses.
Estimates from OLS models. The unit of analysis is the disease level. In Panel A, we report estimates from the full sample of diseases. For each human disease, we identify the highest DisGeNET score among all genes discovered during the exploration period (i.e., before 2000). In Panel B, we report split-sample results that test how our results vary between diseases with more or less competition. Columns (1), (3), and (5) present results for diseases with a bottom-quartile number of principal investigators during the exploration window, while Columns (2), (4), and (6) show results for those with a top-quartile number. We categorize the maximum scores as follows: scores below the 60$^{\text{th}}$ percentile are labeled L, those between the 60$^{\text{th}}$ and 90$^{\text{th}}$ percentiles as M, and those above the 90$^{\text{th}}$ percentile as H. *High-Value Gene*: 0/1=1 if any H candidate was discovered for the disease. *New Genes/Papers*= the number of new genes explored per scientific publication in the years following the exploration period. *Years From 1980*= the number of years until the first H candidate is discovered. In all models, diseases in category L serve as the reference group. We include disease-class fixed effects and control for the number of publications post-2000. See text for more details.

Panel A: Genes with a Mouse Orthologs are Explored Earlier



Panel B: First-Stage Evidence for the Instrumental Variable

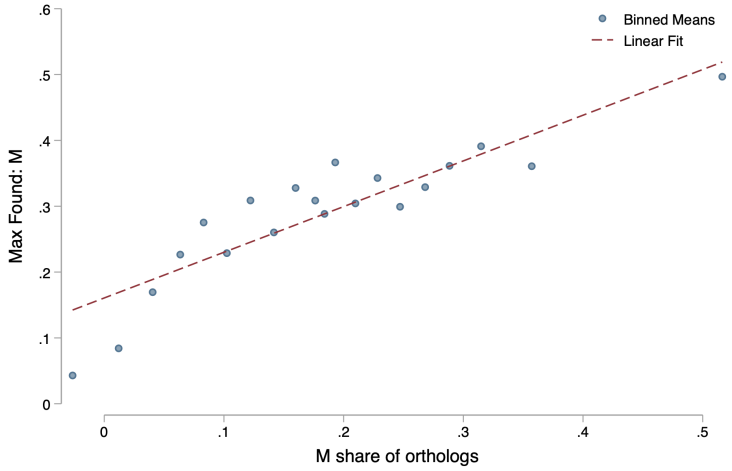


Figure 4: Visual Evidence for Our Instrumental Variable Strategy.

Note: Panel A provides evidence at the gene level that early research tends to focus on genes with mouse orthologs. Each chart shows OLS estimates and 95% confidence intervals estimated from a regression. *First year*= the first year a study exploring a given gene is published. *Explored before 2000*: 0/1=1 if the gene was explored before the year 2000 for at least one disease. Panel B provides a binscatter of the first stage of our disease-level instrumental variable in Table 5. *M Share of Orthologs*: share of orthologous genes (i.e., those with a mouse ortholog) that fall into the M category for each disease. *(Max Found: M)*: 0/1=1 if the maximum DisGeNET score found during the exploration period is classified as M. See text for more details

Table 5: Instrumental Variable Evidence from Human-Mouse Gene Orthologs.

	First Stage	Second Stage		
	Max Found: M	High-Value Gene (0/1)	New Genes/Papers	Years From 1980
	(1)	(2)	(3)	(4)
M Share of Orthologs	0.694***			
	(0.0559)			
Max Found: M		-0.600***	-0.847***	15.93***
		(0.0567)	(0.197)	(2.132)
F-Statistic (First Stage)	154.12			
Disease Class FE	Yes	Yes	Yes	Yes
Count of Publications	Yes	Yes	Yes	Yes
N	4757	4757	4757	4757

Note: † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease-class level in parentheses. We report the effective first-stage F statistic from Olea and Pflueger (2013). Estimates from 2SLS models. The sample is at the disease level. For each human disease, we identify the highest DisGeNET score among all genes discovered during the exploration period (i.e., before 2000). To construct our instrument, we calculate the share of each disease's orthologous gene candidates (i.e., those with a mouse ortholog) that fall into the M category. We categorize the maximum scores as follows: scores below the 60$^{\text{th}}$ percentile are labeled L, those between the 60$^{\text{th}}$ and 90$^{\text{th}}$ percentiles as M, and those above the 90$^{\text{th}}$ percentile as H. *High-Value Gene*: 0/1=1 if any H candidate was discovered for the disease. *New Genes/Papers*= the number of new genes explored per scientific publication in the years following the exploration period. *Years From 1980*= the number of years until the first H candidate is discovered. In all models, diseases in categories L and H serve as the reference group. We include disease-class fixed effects and control for the number of publications post-2000. See text for more details.

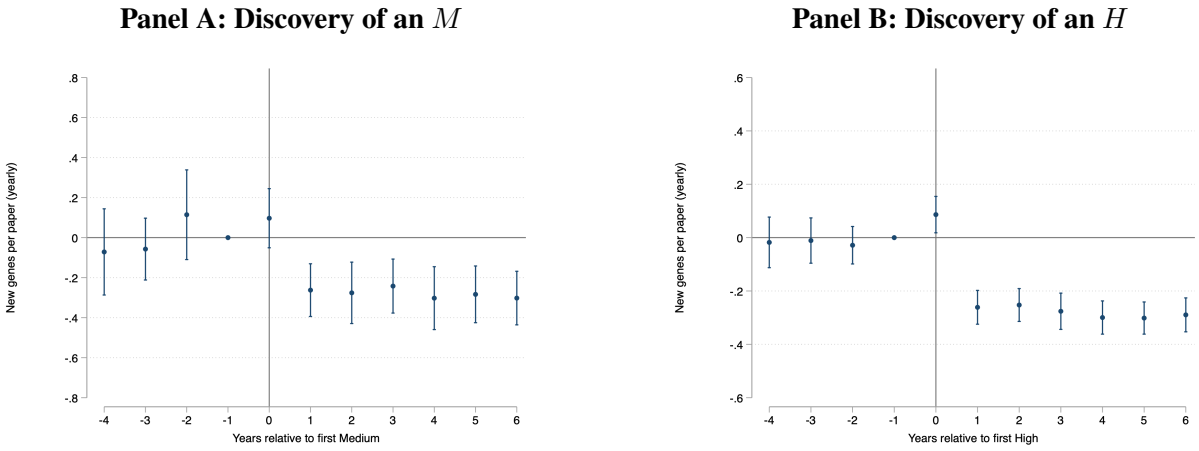


Figure 5: Dynamic Effects of the Discovery of an M or H Genetic Target on Exploration.

Note: Panel A plots OLS estimates and 95% confidence intervals from an event study design that explores how genetic exploration in each disease evolves after the discovery of the first medium-value genetic target. Panel B plots analogous estimates for the discovery of the first high-value genetic target. For each human disease, we classify DisGeNET scores below the 60$^{\text{th}}$ percentile as a "low" gene discovery, scores between the 60$^{\text{th}}$ and 90$^{\text{th}}$ percentile as a "medium" gene discovery, and scores above the 90$^{\text{th}}$ percentile as a "high" (or breakthrough) gene discovery. Standard errors are clustered at the disease class level. See text for more details.

The Streetlight Effect in Data-Driven Exploration

Online Appendix: Additional Results

Johannes Hoelzemann
University of Vienna

Gustavo Manso
UC Berkeley

Abhishek Nagaraj
UC Berkeley & NBER

Matteo Tranchero
The Wharton School

May 17, 2025

A Experimental Results: Additional Details

A.1 Logistics of the Experiment

Figure A.1 summarizes how our experimental sessions unfolded. When participants join, they are assigned either to a data or to a no-data condition.[10] The experiment begins when a total of ten players are assigned to the same experimental set. Then, from each of these experimental sets, two groups of five people are randomly drawn to play the first five rounds (what we labeled as "block"). At the end of the block, the composition of the two groups is randomly reshuffled, and a second block of five rounds is played. This procedure is repeated a total of four times so that each player ends up playing exactly twenty rounds. The order of blocks seen by participants in different experimental sessions is random. The gem types change each round according to a pre-recorded script generated stochastically so that the actual gems and their values each round are effectively random for the player. Similarly, the payoffs and the specific order in which specific gems are revealed in the treatment condition is generated by a random script before the experiment begins.

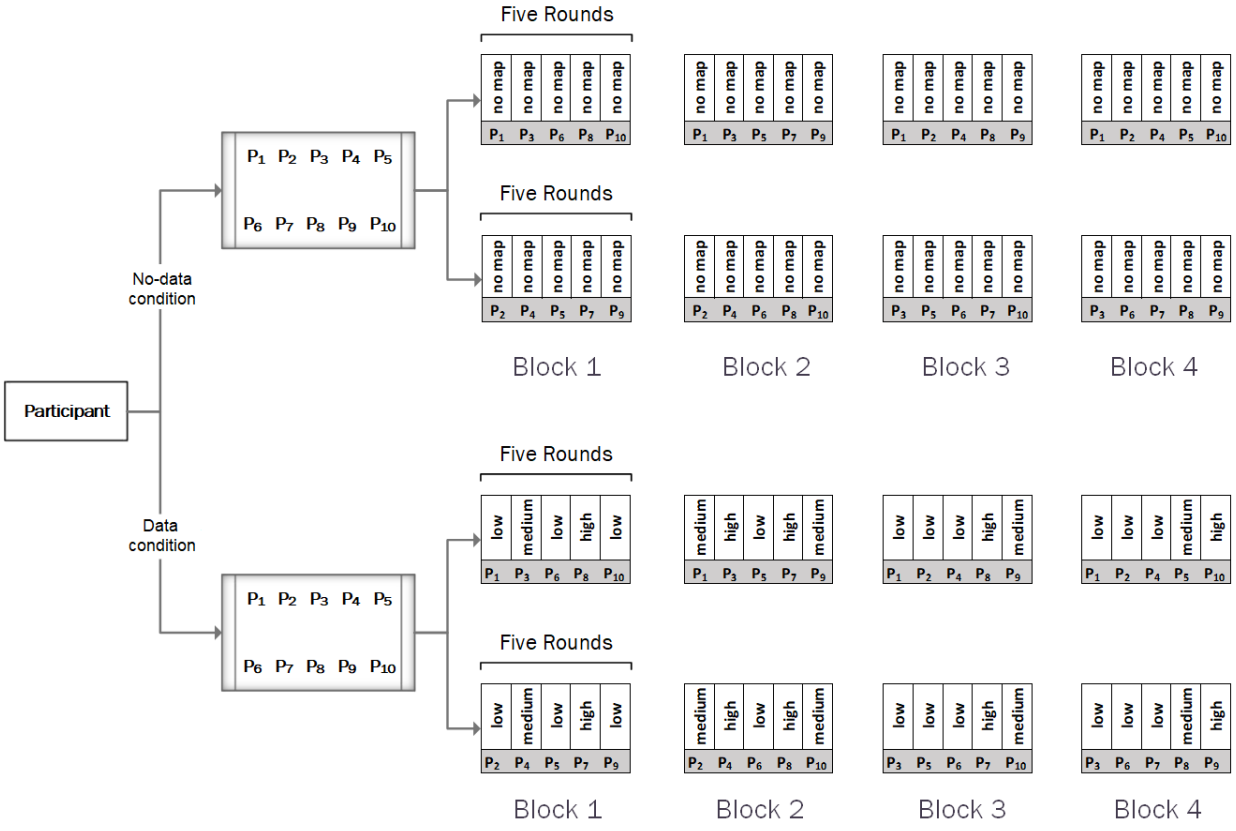


Figure A.1: Flowchart of the Experimental Setup.

Note: This figure provides an overview of the experiment for one actual session that took place in December 2024.

[10] Not in every experimental session there was a no data condition, in which case the players would be randomly split (and then reshuffled) across two distinct data conditions.

Table A.1 presents the descriptive statistics for the experimental data of the main experiment, in which payoffs are non-rivalrous. Data are shown separately by treatment condition. Overall, the table already shows our main results in terms of payoffs, exploration, and discovery.

Table A.1: Descriptive Statistics of the Experimental Data.

	N	Mean	SD	Median	Min	Max
Group Payoff (Share)						
Low	80	75.49	10.87	75.11	49	100
Medium	160	67.64	17.01	66.67	30	100
High	80	98.84	4.11	100.00	71	100
No Data	160	73.27	11.82	72.73	36	100
I(Group found max)						
Low	52	100.00	0.00	100.00	100	100
Medium	112	45.54	50.02	0.00	0	100
High	80	100.00	0.00	100.00	100	100
No Data	120	100.00	0.00	100.00	100	100
Mountains Explored (Share)						
Low	80	89.38	13.65	100.00	50	100
Medium	160	45.31	29.16	50.00	0	100
High	80	3.44	11.76	0.00	0	75
No Data	160	83.25	15.36	80.00	40	100

Note: The table presents descriptive statistics on the 120 participants in the 480 rounds of the experiment with the non-rivalry condition. *Group payoff (Share)*= sum of payoffs as a share of the maximum possible payoff possible in each round; *I(Group found max)*:0/1=1 if the location of the maximum was found by at least one participant in the round; *Group Exploration (Share)*= share of unknown mountains explored in the round.

A.2 Additional Figures and Tables

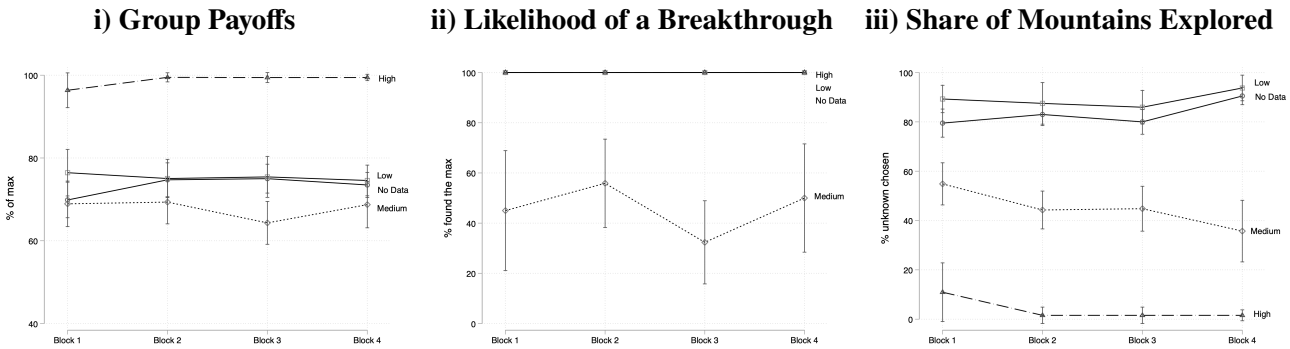
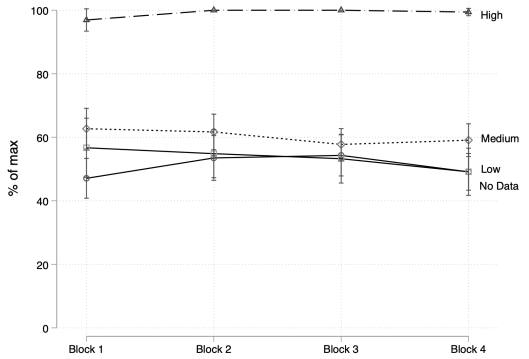


Figure A.2: Outcomes Over Time by Experimental Condition.

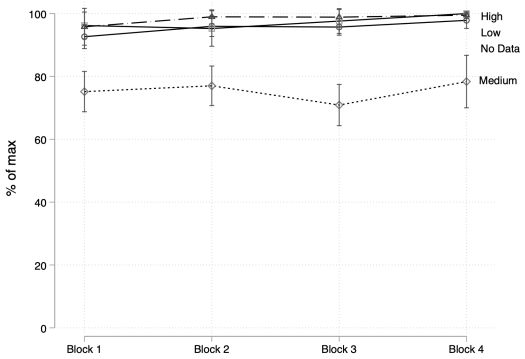
Note: The figures show the impact of data on group outcomes as the experimental session progresses, separately for each block of five rounds. Figure (i) shows the average group payoffs divided by experimental condition. Payoffs are reported as a share of the maximum available in each round. Figure (ii) shows the share of rounds where the maximum was uncovered. Figure (iii) shows the average share of unmapped mountains chosen.

Panel A: Group Payoffs

(i) Payoffs in Period 1

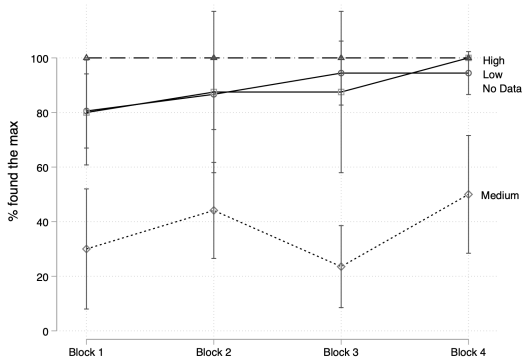


(ii) Payoffs in Period 2



Panel B: Likelihood of a Breakthroughs

(iii) Likelihood of Breakthrough in Period 1



(iv) Likelihood of Breakthrough in Period 2

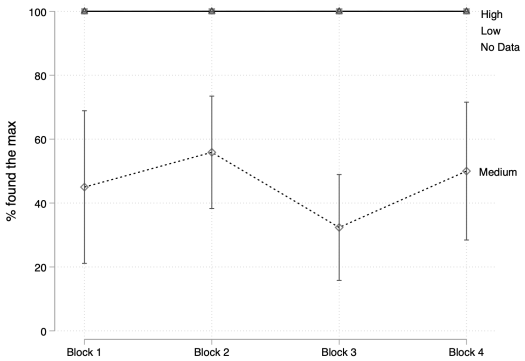


Figure A.3: Outcomes over Time and by Period of the Game.

Note: Panel A reports the experimental results on group payoffs computed as a share of the maximum possible in each period. Figure (i) shows the average group payoffs achieved in period 1 by experimental condition and over time. Figure (ii) shows the average group payoffs achieved in period 2 by experimental condition and over time. Panel B reports the experimental results on the likelihood of a group breakthrough in each round. Figure (iii) shows the share of rounds in which the maximum was uncovered in period 1 by experimental condition and over time. Figure (iv) shows the shows the share of rounds in which the maximum was uncovered in period 2 by experimental condition and over time.

Table A.2: Breaking Down Results by Experimental Period.

	Group Payoff		Group Breakthrough	
	(1) Period 1	(2) Period 2	(3) Period 1	(4) Period 2
High	33.240***	11.281***	8.765**	-1.500
	(0.344)	(0.760)	(2.440)	(4.277)
Low	0.537	1.008	-1.818	-2.045
	(0.338)	(1.094)	(4.764)	(3.631)
Medium	4.837***	-7.969***	-54.948***	-56.338***
	(0.276)	(0.634)	(3.725)	(5.057)
Session FE	Yes	Yes	Yes	Yes
Block order FE	Yes	Yes	Yes	Yes
Payoff structure FE	Yes	Yes	Yes	Yes
Observations	480	480	364	364

Note: † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the session level in parentheses . The sample is at the group-period level (480 rounds). In Columns 3 and 4, the sample only includes rounds that contained at least one diamond (364 rounds). *Group payoff* = group-level period payoffs in Euro; *Group Breakthrough*:0/1=1 if the maximum was found by at least one participant in the period. The excluded category is the control condition without data. See text for more details.

Table A.3: Risk Aversion and Decision Not to Choose the Known Outcome in Period 1 when Medium Is Revealed.

	I(Didn't Choose Medium in Period 1)		
	(1)	(2)	(3)
Risk aversion (std)	-0.088**		
	(0.023)		
Top quartile risk aversion		-0.133*	
		(0.049)	
Bottom quartile risk aversion			0.169*
			(0.050)
Session FE	Yes	Yes	Yes
Block order FE	Yes	Yes	Yes
Payoff structure FE	Yes	Yes	Yes
Round order FE	Yes	Yes	Yes
Observations	800	800	800

Note: † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the session level in parentheses . Round-participant level observations, estimates from OLS models. The sample includes all the individual observations for the 160 rounds where the medium value was revealed. *I(Didn't Choose Medium in Period 1)*:0/1=1 if the player did not choose the medium value in period 1. *Risk aversion*= standardized measure of individual risk aversion (Holt and Laury, 2002); *Top quartile risk aversion*:0/1=1 if the participant is in the top quartile of the risk aversion distribution in our sample; *Bottom quartile risk aversion*:0/1=1 if the participant is in the bottom quartile of the risk aversion distribution in our sample.

Table A.4: Correlates of the Decision not to Choose the Known Outcome in Period 1 when Medium Is Revealed.

	I(Didn't Choose Medium in Period 1)			
	(1)	(2)	(3)	(4)
English native	-0.064			
	(0.072)			
Wrong quizzes (std)		0.048		
		(0.030)		
Round number			-0.002	
			(0.006)	
Order of choice				0.009
				(0.015)
Session FE	Yes	Yes	Yes	Yes
Block order FE	Yes	Yes	Yes	Yes
Payoff structure FE	Yes	Yes	Yes	Yes
Round order FE	Yes	Yes	Yes	No
Observations	800	800	800	800

Note: † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the session level in parentheses . Round-participant level observations, estimates from OLS models. The sample includes all the individual observations for the 160 rounds where the medium value was revealed. *I(Didn't Choose Medium in Period 1)*:0/1=1 if the player did not choose the medium value in period 1. *English native*:0/1=1 if the participant is a native English speaker based on her reported nationality; *Wrong quizzes*= standardized number of wrong answers to the initial comprehension test; *Round number*= progressive order in which the rounds were played in the experimental session; *Order of choice*= random sequential order in which the player chose in that round.

B The Genetic Roots of Human Diseases: Additional Details

B.1 Scientific Background

Genetics is the branch of biology that studies genes, heredity, and variation in living organisms. Genes are segments of DNA (deoxyribonucleic acid) that contain the information necessary for living organisms' development, functioning, and reproduction. In practice, each gene is a portion of DNA that contains instructions for building one or more products, such as proteins, which are the fundamental constituents of an organism. Genes often acquire mutations (or variants) in their sequence, most of which are harmless. However, some mutations can lead the gene to alter its behavior and affect phenotypic traits, sometimes with significant consequences and the emergence of severe health conditions. Discovering which mutations are responsible for specific human diseases is thus a first-order priority since genes associated with a condition can often be used as drug targets (Nelson et al., 2015). When a drug molecule binds to its genetic target, it can modify its functioning, favorably affecting the outcome of a disease. Therefore, knowing the genetic roots of diseases has important practical consequences in the design of pharmaceutical drugs.

Diseases caused by single gene mutations are called Mendelian disorders, but such diseases are typically rare. Most common human diseases have a polygenic nature, meaning they are not due to a single genetic factor but rather by mutations in many genes. This class of diseases is called complex and genetic mutations may increase the risk of developing the condition without being either necessary or sufficient on their own. Despite often clustering in families, polygenic disorders do not have a predictable inheritance pattern because convoluted interactions between genes and environmental factors determine them. This means that scientists need to search through the over 19,000 protein-coding genes to find the mutations involved in each of the thousands of polygenic diseases (Tranchero, 2025).

Researchers have noted that even after the completion of the Human Genome Project, most scientists continue to investigate the same small number of genes (Stoeger et al., 2018). Gates et al. (2021) report that 1% of genes still receive 22% of all gene-related publications, helping to explain why current treatments exploit only around 10% of the potentially druggable targets. This situation is probably suboptimal since our chances of finding a cure for polygenic diseases would benefit from exploring a larger number of genes (Edwards et al., 2011) and several understudied genes showing high promise have been identified (Nguyen et al., 2017; Stoeger et al., 2018). Interestingly, despite much debate on this extreme concentration of attention on a small number of theoretically well-known genes, we still lack an explanation for its drivers. Some scholars have attributed it to scientists' preference for genes with past data that permit the formulation of functional hypotheses (Haynes et al., 2018), akin to what we characterized as a streetlight effect in this paper.

B.2 Data Description

DisGeNET. Our main data source is DisGeNET (v7.0), which is considered a complete repository of scientific results linking human diseases to their genetic causes (Piñero et al., 2020). This database

aggregates all novel gene-disease combinations studied by publications indexed in PubMed. The information is harvested from specialized sources, including curated datasets such as ClinVar, UniProt, and Orphanet.[11] In addition, DisGeNET complements these data with information extracted from the scientific literature indexed in PubMed using text-mining approaches. Our starting data are at the gene-disease-paper level, because for each gene-disease pair we observe both the publication that introduced it and the list of all follow-up articles that investigated it.

Genes. Each gene in the database is identified by a unique ID from Entrez Gene, a gene-centric resource maintained by the National Center for Biotechnology Information (NCBI). These identifiers are species-specific, meaning the ID assigned to a human gene differs from that of its homolog in another species. DisGeNET includes only data from studies on human genes and compiles the Entrez Gene ID for each gene examined in PubMed-indexed papers. We further restrict our sample to protein-coding genes, given their central role in the drug discovery process (Nelson et al., 2015).

Diseases. Disease entries in DisGeNET are annotated using vocabulary from the Unified Medical Language System (UMLS), a set of crosswalks that bring together many health and biomedical vocabularies and standards to enable interoperability between databases. DisGeNET compiles the UMLS ID of each disease studied by papers in PubMed. Since we focus on human diseases, we keep any entries that map to the following UMLS semantic types: disease or syndrome; neoplastic process; acquired abnormality; anatomical abnormality; congenital abnormality; and mental or behavioral dysfunction. Using the UMLS ID, we also obtain disease relations from Kehoe and Torvik (2019), which contains all pairwise relationships in the Medical Subject Headings vocabulary (MeSH) hierarchy.

B.3 DisGeNET Score

DisGeNET is designed to help researchers in both academia and industry prioritize promising genetic targets based on existing knowledge. To support this goal, it provides a synthetic DisGeNET Score for each gene–disease pair. The Score ranges from 0 to 1, with higher values indicating combinations that are more scientifically robust and therapeutically promising. It incorporates both the curation and reliability of the sources supporting a given association, as well as the number of publications that have studied it. In practice, the Score reflects how well-established a gene target is in the current literature. In the version used in this paper (v7.0), the score offers a parsimonious way to assess the scientific strength of any given gene–disease pair as of 2020.

In particular, the raw DisGeNET score is build with the following formula:

$$\text{DisGeNET score of gene i for disease j} = C_{i,j} + M_{i,j} + I_{i,j} + L_{i,j}$$

The first component $C_{i,j}$ summarizes the evidence from curated sources reporting gene-disease com-

[11] For the complete list of sources aggregated by DisGeNET, see https://www.disgenet.org/dbinfo.

bination $< i, j >$:

$$C_{i,j} = \begin{cases} 0.6 & \text{if } N_{sources_c} > 2 \\ 0.5 & \text{if } N_{sources_c} = 2 \\ 0.3 & \text{if } N_{sources_c} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{B.1}$$

where $N_{sources_c}$ is the number of curated sources supporting a gene-disease association, including CGI, ClinGen, Genomics England, CTD, PsyGeNET, Orphanet, and UniProt.

The second component $M_{i,j}$ summarizes the evidence from experiments using mice models reporting gene-disease combination $< i, j >$:

$$M_{i,j} = \begin{cases} 0.2 & \text{if } N_{sources_m} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{B.2}$$

where $N_{sources_m}$ is the number of sources using the lab rat or lab mouse from RGD, MGD, and CTD.

The third component $I_{i,j}$ summarizes the evidence inferred from experiments on gene-disease combination $< i, j >$:

$$I_{i,j} = \begin{cases} 0.1 & \text{if } N_{sources_i} > 1 \\ 0 & \text{otherwise} \end{cases} \tag{B.3}$$

where $N_{sources_i}$ is the number of sources from HPO, CLINVAR, GWAS Catalog, and GWASDB.

Finally, the component $L_{i,j}$ summarizes the evidence mined from the literature about gene-disease combination $< i, j >$:

$$L_{i,j} = \begin{cases} 0.1 & \text{if } N_{publications} > 9 \\ \text{N}_{publications} \cdot 0.01 & \text{if } N_{publications} \leq 9 \end{cases} \tag{B.4}$$

where $N_{publications}$ is the number of publications supporting a gene-disease association as mined by LHGDN and BEFREE.

The DisGeNET Score has strong face validity and has been thoroughly validated in prior research (Piñero et al., 2020). Because it is designed to capture the biological importance of a gene-disease pair, we should expect higher-scoring associations to be linked to more downstream pharmaceutical development—such as clinical citations, granted patents, and approved drugs. To test this, we regress the raw DisGeNET score on each of these real-world outcomes. The results, presented in Appendix Table B.1, show that higher scores are associated with significantly greater levels of clinical citations, patenting activity, and drug development.

Table B.1: Associations Between DisGeNET Scores and Real-World Pharmaceutical Outcomes

	Clinical Citations		Granted Patents		Approved Drugs	
	(1) Count (#)	(2) Has Any (0/1)	(3) Count (#)	(4) Has Any (0/1)	(5) Count (#)	(6) Has Any (0/1)
DisGeNET Score	51.428***	0.939***	1.669***	0.153***	0.085***	0.022***
	(2.302)	(0.005)	(0.070)	(0.003)	(0.007)	(0.001)
N	810,377	810,377	810,377	810,377	810,377	810,377

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust standard errors in parentheses. Estimates from OLS models. The sample is at the gene-disease level. We correlate the raw DisGeNET score with real-world measures of clinical impact. *Count Clinical* = total clinical articles based on a gene-disease pair. *Granted Patents* = count of USPTO granted patents for inventions leveraging a given gene as a drug target for a given disease. *Count Drugs* = count of FDA-approved drugs leveraging a given gene as a drug target for a given disease. Models (1), (3), and (5) use count variables, while Models (2), (4), and (6) use corresponding indicator versions.

In our main specification, we convert the raw DisGeNET scores into percentile ranks, and consider any score below the 60$^{\text{th}}$ percentile as a low payoff, between the 60$^{\text{th}}$ and 90$^{\text{th}}$ percentile as a medium payoff, and above the 90$^{\text{th}}$ percentile as a high payoff, respectively. We now verify that these score categories correspond to meaningful differences in real-world outcomes. In Appendix Figure B.1, we plot our three outcome metrics by score category. We find that clinical citations, granted patents, and approved drugs are all increasing in score category, suggesting that our score thresholds do capture substantive differences in impact.

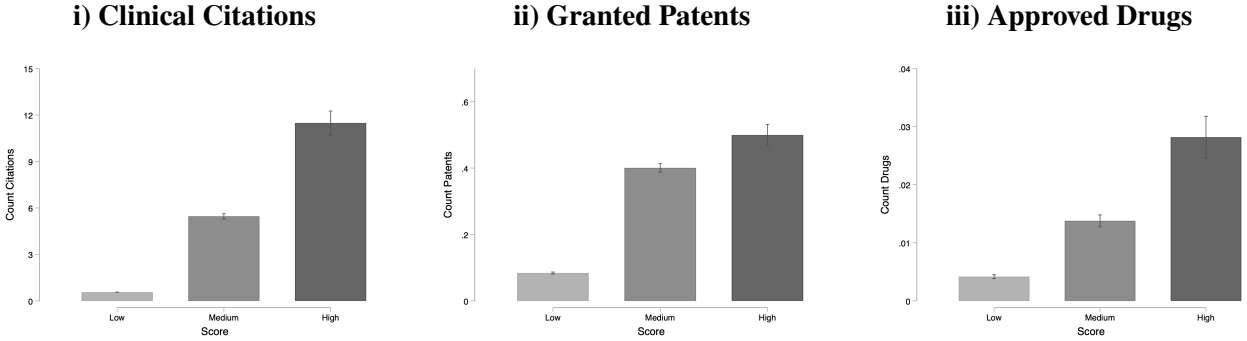


Figure B.1: Relationship Between Score Categories and Real-World Outcomes.

Note: This figure shows the relationship between our categorization of raw DisGeNET scores (Low, Medium, High) and real-world innovation outcomes. Panel (i) shows the average number of clinical citations on a gene-disease pair (data as of 2024). Panel (ii) shows the average number of granted patents targeting a gene-disease pair (data as of 2023). Panel (iii) shows the average number of approved drugs targeting a gene-disease pair (data as of 2023).

One potential limitation is that the DisGeNET score is time-invariant, since it is calculated ex post across the full sample period, incorporating all available evidence up to the time when our data were collected (2020). While this gives us the best available evidence of association strength at the time of analysis, it does raise concerns about the variability of the score over time. In particular, such a bias could arise if scientists initially pursued a genetic target thinking it was high value (H), only for it to be revised downward into the M range at a later moment. Alternatively, researchers may have believed an M-value gene-disease pair had the potential to become an H with further investigation

and behaved accordingly. To assess the validity of this concern, we compare the current DisGeNET scores to those from the first release in 2015 (version 1). As shown in Figure B.2, the ordinal rankings of gene-disease pairs are largely preserved over time, showing a remarkable stability of the DisGeNET scores. To further ease this concern, we replicate our main analysis using the 2015 scores. If the assumption of time-invariance is reasonable, then we should obtain the same results. As reported in Table B.2, the results are consistent with our main analysis.

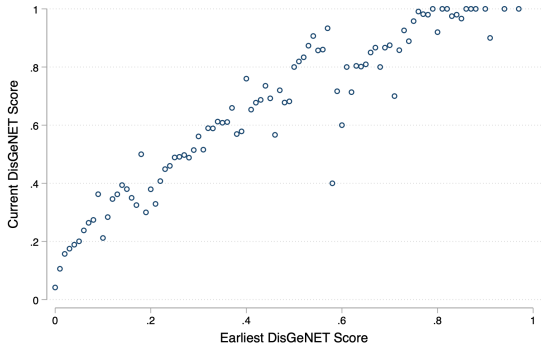


Figure B.2: Stability of DisGeNET Scores Over Time

Note: This figure presents a binned scatterplot comparing the earliest available DisGeNET scores (version 1, released in 2015) to those from the release used in our main analysis (version 7, released in 2020).

Table B.2: Analysis Based on DisGeNET v1 Scores (2015)

	Group Breakthrough	Group Exploration	Group Delay
	(1) High-Value Gene (0/1)	(2) New Genes/Papers	(3) Years From 1980
Max Found: M	-0.049†	-0.151***	0.939*
	(0.029)	(0.029)	(0.441)
Max Found: H	0.739***	-0.273***	-23.229***
	(0.031)	(0.030)	(0.595)
Disease Class FE	Yes	Yes	Yes
Count of Publications	Yes	Yes	Yes
N	3261	3261	3261

Note: $\dagger$ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease-class level in parentheses.
Estimates from OLS models. The sample is at the disease-level. This table replicates our baseline specification but uses the earliest version of DisGeNET scores (released in 2015). For each human disease, we determine the highest DisGeNET score for any gene identified during the exploration period (i.e., pre-2000). We classify maximum scores below the 60$^{\text{th}}$ percentile as L, scores between the 60$^{\text{th}}$ and 90$^{\text{th}}$ percentile as M, and scores above the 90$^{\text{th}}$ percentile as H. *High-Value Gene*: 0/1=1 if any H candidate was discovered for the disease. *New Genes/Papers*= the number of new genes explored per scientific publication in the years following the exploration period. *Years From 1980*= the number of years until the first H candidate is discovered. In all models, diseases in category L serve as the reference group. We include disease-class fixed effects and control for the number of publications post-2000. See text for more details.

C Additional Figures and Tables

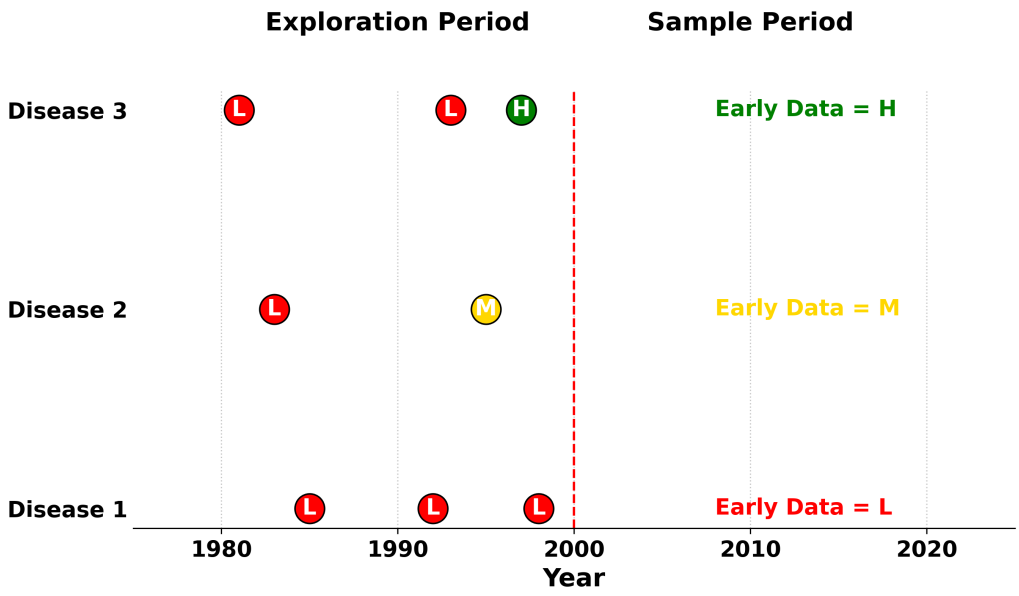


Figure C.1: Illustration of the Empirical Setup.

Note: This figure presents a stylized depiction of our approach to translate the theoretical framework to the disease-level data in our sample. For each human disease, we record every gene identified during the early exploration period (i.e., pre-2000). We classify scores below the 60$^{\text{th}}$ percentile as L (red), scores between the 60$^{\text{th}}$ and 90$^{\text{th}}$ percentile as M (yellow), and scores above the 90$^{\text{th}}$ percentile as H (green). The highest-scoring genetic target for each disease is then used to classify the nature of early data available to scientists. In this stylized representation, scientists identified three gene-disease pairs with L scores for Disease 1, which means we classify its early data as L. For Disease 2, scientists found one L and one M, resulting in a classification of M. For Disease 3, two L scores and one H score were uncovered, leading to a classification of H. See text for further details.

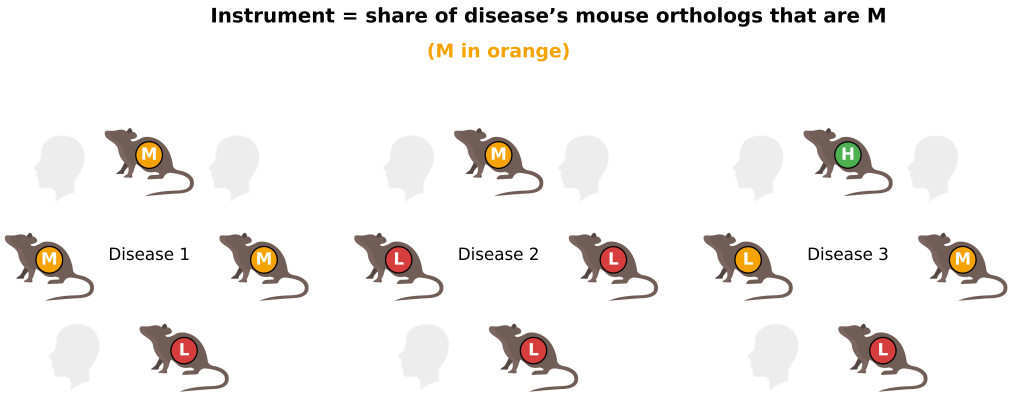


Figure C.2: Illustration of the Instrumental Variable Strategy.

Note: This figure presents a stylized depiction of our IV approach, which relies on gene orthology (i.e., when genes in different species descend from a common ancestor, largely retaining the same biological function). For each disease, we only consider the gene candidates that have an ortholog in a mouse. We then measure the share of these orthologous genes that are classified as medium-value (M) candidates. In the example above, the M share for Disease 1 would be 75%, while for Disease 2 it would be 25%. See text for further details.

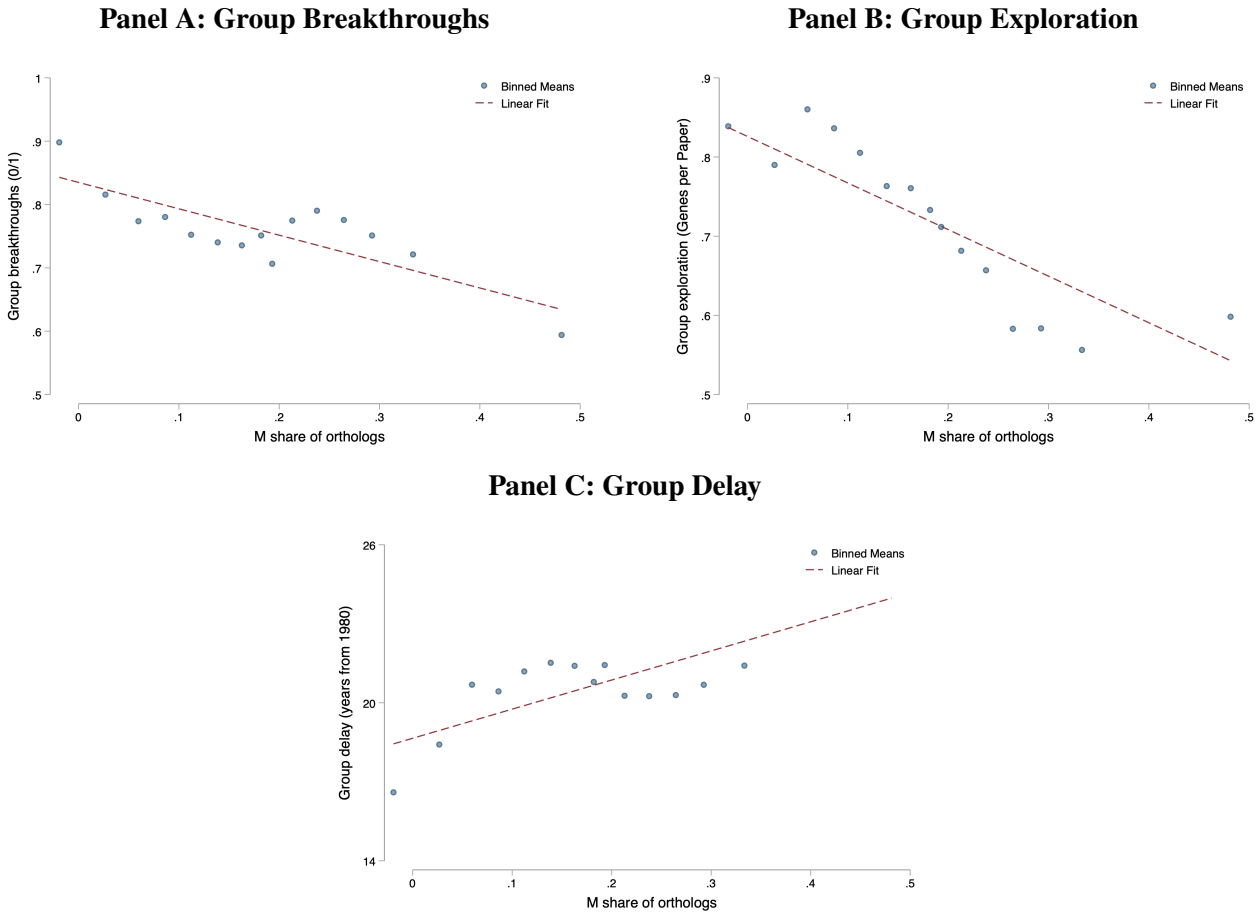
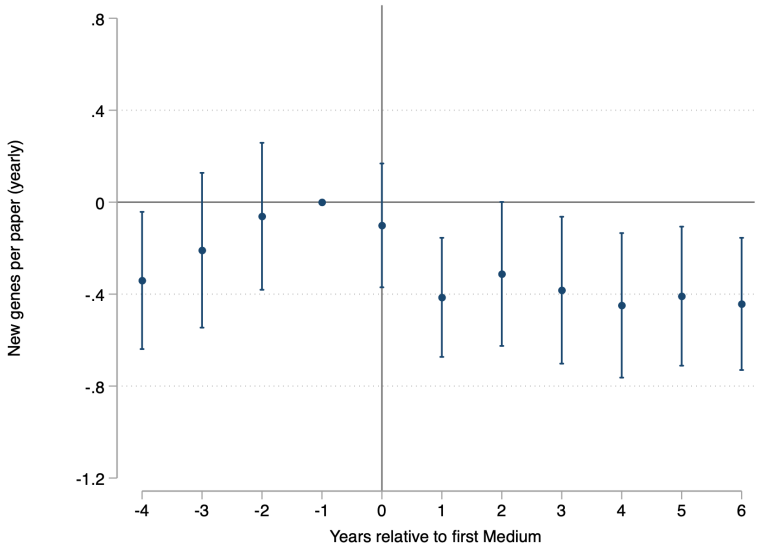


Figure C.3: Reduced-Form Evidence for the Streetlight Effect in the Search for Genetic Candidates.

Note: This figure shows binned scatterplots for each of our dependent variables against our instrumental variable, defined as the share of each disease's orthologous genes that fall into the M category. Panel (i) shows the impact of the instrument on the likelihood of finding any breakthrough during the sample period. Panel (ii) shows the impact of the instrument on the number of new genes explored per publication in the years following the exploration window. Panel (iii) shows the impact of the instrument on the delay in discovering a breakthrough, defined as the years elapsed from 1980 (the first year of our panel). We include controls for disease class and the number of publications post-2000. See text for more details.

Panel A: Keeping Sibling and Parent Diseases



Panel B: Keeping only Sibling Diseases

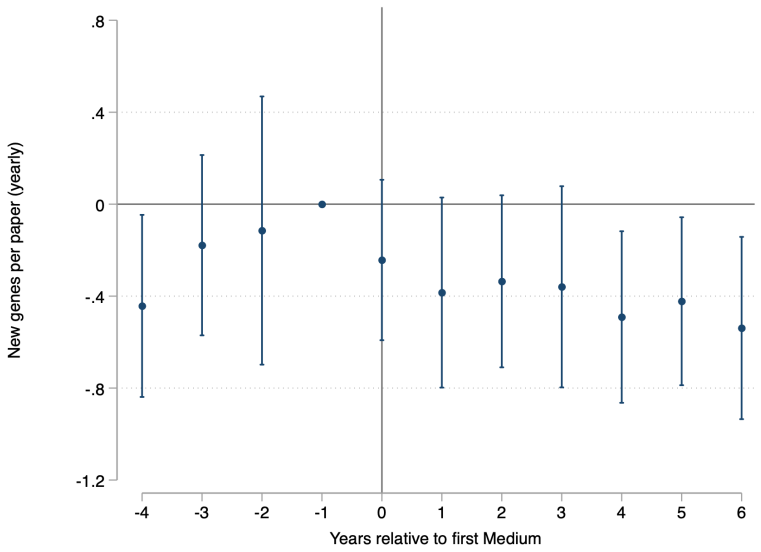


Figure C.4: Considering only Diseases Genetically Related to a Disease with a Breakthrough.

Note: This figure replicates the event study of figure 5 but only considers diseases that are genetically related to a disease with a known breakthrough (genetic discoveries with scores above the 90th percentile of DisGeNET scores). We obtain genetic relations from the Medical Subject Headings vocabulary (MeSH). In Panel A, we keep both sibling diseases (i.e., diseases sharing the same parent MeSH code) and parent diseases (i.e., diseases one level up in the MeSH tree) of diseases with a breakthrough. In Panel B, we keep only sibling diseases (i.e., diseases sharing the same parent MeSH code) of diseases with a breakthrough. This figure plots OLS estimates and 95% confidence intervals from an event study design that explores how genetic exploration in each disease evolves in the years before and after the discovery of the first medium-value genetic association. Standard errors are clustered at the disease class level. See text for more details.

Table C.1: Association Between Genes Having Mouse Orthologs and Their Appearance in the Scientific Literature.

Panel A: Gene Level

	(1) Publication Year	(2) Publication Year
Has Mouse Ortholog (0/1)	-0.798**	-2.559***
	(0.262)	(0.372)
Gene Group FE	No	Yes
N	16,000	10,344

Panel B: Gene-Disease Level

	(1) Publication Year	(2) Publication Year
Has Mouse Ortholog (0/1)	-0.00385***	-0.0122***
	(0.0000985)	(0.000169)
Disease FE	Yes	Yes
Gene Group FE	No	Yes
N	339,136,000	257,764,556

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust standard errors in parentheses.
This table examines whether human genes with orthologous counterparts in the lab mice are explored earlier by scientists. In Panel A, the data is at the gene level. We assess whether genes with an ortholog appear in scientific studies earlier regardless of the disease. In Panel B, the data is at the gene-disease level. We assess whether genes with an ortholog appear in scientific studies earlier for a given disease. We impute a value of 2020 for gene-disease pairs with no recorded publications. In all models, the dependent variable is the first year of publication. In Column (2) of both Panel A and Panel B, we include controls for gene group classification.

Table C.2: Descriptive Statistics at the Disease-Year Level.

	Mean	Median	Sd	Min	Max	N
Maximum Gene Score	39.68	0.00	45.56	0	100	220,760
Found Any High Gene (0/1)	0.28	0.00	0.45	0	1	220,760
Count of Publications	7.37	0.00	82.60	0	10,449	220,760
Count of Genes Discovered	3.25	0.00	15.46	0	685	220,760
New Genes per Paper	0.74	0.67	1.03	0	123	109,002

Note: This table presents descriptive statistics for our disease-year panel. *Maximum Gene Score* = the highest DisGeNET score uncovered each year for a disease. *Found Any High Gene*:0/1=1 if any H was discovered a year for a given disease. *Count of Publications* = the number of publications on the disease in a given year. *Count of Genes Discovered* = the number of genes explored in relation to a disease each year. *New Genes Per Paper* = the number of genes explored per paper for a disease each year.

Table C.3: Disease-Year-Level Analysis of Exploration Dynamics.

	Group Exploration	Group Exploration
	(1) New Genes/Papers	(2) New Genes/Papers
Post M Discovery	-0.180***	
	(0.030)	
Post H Discovery		-0.256***
		(0.022)
Disease FE	Yes	Yes
Year FE	Yes	Yes
Count of Publications	Yes	Yes
N	98,547	97,956

Note: † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease-class level in parentheses.

OLS estimates from differences-in-differences models. The sample is at the disease-year level. We examine how genetic exploration within each disease evolves following the discovery of the first medium-value (Column 1) and high-value (Column 2) genetic target. For each disease, we classify DisGeNET scores below the 60th percentile as "low," scores between the 60th and 90th percentiles as "medium," and scores above the 90th percentile as "high" (or breakthrough) discoveries. *Yearly Genes/Papers*= the number of new genes explored per scientific publication. All models include disease fixed effects and year fixed effects, and control for the annual number of publications. See text for more details.

Table C.4: Robustness to TWFE Weighting Concerns.

	csdid		did_multiplegt		did_imputation	
	(1) Genes/Papers	(2) Genes/Papers	(3) Genes/Papers	(4) Genes/Papers	(5) Genes/Papers	(6) Genes/Papers
Post M Discovery	-0.352***		-0.238**		-0.246***	
	(0.069)		(0.081)		(0.037)	
Post H Discovery		-0.223***		-0.187***		-0.281***
		(0.031)		(0.031)		(0.029)
Disease FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Count of Publications	Yes	Yes	Yes	Yes	Yes	Yes
N	56870	49812	48619	32415	64503	59402

Note: † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease-class level in parentheses.

OLS estimates from differences-in-differences models. The sample is at the disease-year level. We replicate Table C.4 but use alternate estimators that avoid weighting problems associated with TWFE. We implement the csdid command from Callaway and Sant'Anna (2020) in Columns 1-2, the did_multiplegt_dn command from Chaisemartin and D'Haultfoeuille (2024) in Columns 3-4, and the did_imputation command from Borusyak et al. (2021) in Columns 5-6. For each disease, we classify DisGeNET scores below the 60th percentile as "low," scores between the 60th and 90th percentiles as "medium," and scores above the 90th percentile as "high" (or breakthrough) discoveries. *Yearly Genes/Papers*= the number of new genes explored per scientific publication. All models include disease fixed effects and year fixed effects, and control for the annual number of publications. See text for more details.

Table C.5: Considering only Diseases Genetically Related to a Disease with a Breakthrough.

	Siblings and Parents	Siblings Only
	(1) Yearly Genes/Papers	(2) Yearly Genes/Papers
Post M Discovery	-0.251***	-0.287**
	(0.061)	(0.092)
Disease FE	Yes	Yes
Year FE	Yes	Yes
Count of Publications	Yes	Yes
N	40505	24416

Note: † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease-class level in parentheses.

OLS estimates from differences-in-differences models. The sample is at the disease-year level. We replicate Table C.5, but only considers diseases in the sample that are genetically related to a disease with a known breakthrough (genetic discoveries with scores above the 90th percentile of DisGeNET score). We obtain genetic relations from the Medical Subject Headings vocabulary (MeSH). In Column 1, we keep both sibling diseases (i.e., diseases sharing the same parent MeSH code) and parent diseases (i.e., diseases one level up in the MeSH tree) of diseases with a breakthrough. In Column 2, we keep only sibling diseases (i.e., diseases sharing the same parent MeSH code) of diseases with a breakthrough. For each disease, we classify DisGeNET scores below the 60th percentile as "low," scores between the 60th and 90th percentiles as "medium," and scores above the 90th percentile as "high" (or breakthrough) discoveries. *Yearly Genes/Papers*= the number of new genes explored per scientific publication. All models include disease fixed effects and year fixed effects, and control for the annual number of publications. See text for more details.

Table C.6: Sensitivity to Definition of Marginally Explored Diseases.

	Group Breakthrough			Group Exploration			Group Delay		
	(1) >5 Pubs	(2) >15 Pubs	(3) >25 Pubs	(4) >5 Pubs	(5) >15 Pubs	(6) >25 Pubs	(7) >5 Pubs	(8) >15 Pubs	(9) >25 Pubs
Max Found: M	-0.043	-0.121**	-0.162***	-0.205***	-0.131***	-0.109***	0.883*	2.056***	2.773***
	(0.028)	(0.038)	(0.047)	(0.019)	(0.027)	(0.032)	(0.425)	(0.583)	(0.745)
Max Found: H	0.602***	0.477***	0.422***	-0.315***	-0.260***	-0.243***	-21.400***	-19.910***	-19.271***
	(0.037)	(0.045)	(0.056)	(0.029)	(0.033)	(0.038)	(0.594)	(0.729)	(0.909)
Disease Class FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Count of Publications	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	5641	4149	3355	5641	4149	3355	5641	4149	3355

Note: † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease-class level in parentheses.

Estimates from OLS models. The sample is at the disease-level. This table replicates our baseline specification (which removes diseases with less than 10 publications over the sample window) and shows robustness when we keep only diseases with more than 5 publications (Columns (1), (4), and (7)), more than 15 publications (Columns (2), (5), and (8)), and more than 25 publications (Columns (3), (6), and (9)). For each human disease, we determine the highest DisGeNET score for any gene identified during the exploration period (i.e., pre-2000). We classify maximum scores below the 60$^{\text{th}}$ percentile as L, scores between the 60$^{\text{th}}$ and 90$^{\text{th}}$ percentile as M, and scores above the 90$^{\text{th}}$ percentile as H. Columns 1-3 show the impact of early discoveries on the likelihood of finding any breakthrough during the sample period. Columns 4-6 show the impact on the number of new genes explored per scientific publication. Columns 7-9 show the impact of the delay in discovering a breakthrough, defined as the years that elapsed from 1980 (the first year of our panel). In all models, diseases classified under L constitute the excluded category. We include disease-class fixed effects and control for the number of publications post-2000. See text for more details.

Table C.7: Sensitivity to the Exclusion of Outlier Diseases.

	Group Breakthrough	Group Exploration	Group Delay
	(1) High-Value Gene (0/1)	(2) New Genes/Papers	(3) Years From 1980
Max Found: M	-0.105**	-0.129***	1.878***
	(0.033)	(0.023)	(0.514)
Max Found: H	0.503***	-0.200***	-19.465***
	(0.042)	(0.036)	(0.642)
Disease Class FE	Yes	Yes	Yes
Count of Publications	Yes	Yes	Yes
N	4675	4675	4675

Note: † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease-class level in parentheses.

Estimates from OLS models. The sample is at the disease-level. This table replicates our baseline specification but excluding outlier diseases (i.e., those in the top 1% by publications over the sample period). For each human disease, we determine the highest DisGeNET score for any gene identified during the exploration period (i.e., pre-2000). We classify maximum scores below the 60$^{\text{th}}$ percentile as L, scores between the 60$^{\text{th}}$ and 90$^{\text{th}}$ percentile as M, and scores above the 90$^{\text{th}}$ percentile as H. *High-Value Gene*: 0/1=1 if any H candidate was discovered for the disease. *New Genes/Papers*= the number of new genes explored per scientific publication in the years following the exploration period. *Years From 1980*= the number of years until the first H candidate is discovered. In all models, diseases in category L serve as the reference group. We include disease-class fixed effects and control for the number of publications post-2000. See text for more details.

Table C.8: Alternative Definitions of Low and Medium-Value Genes.

	Group Breakthrough			Group Exploration			Group Delay		
	(1) 60^{th} P	(2) 70^{th} P	(3) 80^{th} P	(4) 60^{th} P	(5) 70^{th} P	(6) 80^{th} P	(7) 60^{th} P	(8) 70^{th} P	(9) 80^{th} P
Max Found: M	-0.105**	-0.072**	-0.082***	-0.144***	-0.180***	-0.259***	1.743***	1.200**	1.303***
	(0.033)	(0.025)	(0.024)	(0.023)	(0.020)	(0.019)	(0.519)	(0.425)	(0.391)
Max Found: H	0.514***	0.549***	0.555***	-0.261***	-0.261***	-0.269***	-20.371***	-20.961***	-21.093***
	(0.042)	(0.033)	(0.031)	(0.028)	(0.027)	(0.029)	(0.692)	(0.589)	(0.537)
Disease Class FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Count of Publications	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	4760	4760	4760	4760	4760	4760	4760	4760	4760

Note: † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease-class level in parentheses.

Estimates from OLS models. The sample is at the disease-level. This table replicates our baseline specification but varies the cutoff between a low and medium-value genetic association. In our baseline, we adopt the 60th percentile to separate medium and high scores. We test the baseline (Columns (1), (4), and (7)), the 70th percentile (Columns (2), (5), and (8)), and the 80th percentile (Columns (3), (6), and (9)) instead. For each model, we hold the cutoff between a medium gene score and a high gene score fixed at the 90th percentile. For each human disease, we determine the highest DisGeNET score for any gene identified during the exploration period (i.e., pre-2000). Columns 1-3 show the impact of early discoveries on the likelihood of finding any breakthrough during the sample period. Columns 4-6 show the impact on the number of new genes explored per scientific publication in the years following the exploration period. Columns 7-9 show the impact on the delay in discovering a breakthrough, defined as the years that elapsed from 1980 (the first year of our panel). In all models, diseases classified under L constitute the excluded category. We include disease-class fixed effects and control for the number of publications post-2000. See text for more details.

Table C.9: Alternative Definitions of Medium and High-Value Genes.

	Group Breakthrough			Group Exploration			Group Delay		
	(1) 90^{th} P	(2) 95^{th} P	(3) 100^{th} P	(4) 90^{th} P	(5) 95^{th} P	(6) 100^{th} P	(7) 90^{th} P	(8) 95^{th} P	(9) 100^{th} P
Max Found: M	-0.105**	-0.023	-0.053*	-0.144***	-0.110***	-0.142***	1.743***	0.588	1.048**
	(0.033)	(0.026)	(0.022)	(0.023)	(0.023)	(0.023)	(0.519)	(0.388)	(0.350)
Max Found: H	0.514***	0.667***	0.802***	-0.261***	-0.364***	-0.466***	-20.371***	-21.905***	-23.137***
	(0.042)	(0.036)	(0.034)	(0.028)	(0.027)	(0.027)	(0.692)	(0.652)	(0.622)
Disease Class FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Count of Publications	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	4760	4760	4760	4760	4760	4760	4760	4760	4760

Note: † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease-class level in parentheses.

Estimates from OLS models. The sample is at the disease-level. This table replicates our baseline specification but varies the cutoff between a medium and high-value genetic association. In our baseline, we adopt the 90th percentile to separate medium and high scores. We test the baseline (Columns (1), (4), and (7)), the 95th percentile (Columns (2), (5), and (8)), and the 100th percentile (Columns (3), (6), and (9)). For each model, we hold the cutoff between a low gene score and a medium gene score fixed at the 60th percentile (our baseline). Columns 1-3 show the impact of early discoveries on the likelihood of finding any breakthrough during the sample period. Columns 4-6 show the impact on the number of new genes explored per scientific publication in the years following the exploration period. Columns 7-9 show the impact on the delay in discovering a breakthrough, defined as the years that elapsed from 1980 (the first year of our panel). In all models, diseases classified under L constitute the excluded category. We include disease-class fixed effects and control for the number of publications post-2000. See text for more details.

Table C.10: Alternative Definitions of the Early Exploration Period.

	Group Breakthrough			Group Exploration			Group Delay		
	(1) <1995	(2) <2000	(3) <2005	(4) <1995	(5) <2000	(6)<2005	(7) <1995	(8) <2000	(9)<2005
Max Found: M	-0.024	-0.105**	-0.100**	-0.150***	-0.144***	-0.223***	0.646	1.743***	1.172**
	(0.036)	(0.033)	(0.036)	(0.021)	(0.023)	(0.028)	(0.663)	(0.519)	(0.435)
Max Found: H	0.435***	0.514***	0.656***	-0.223***	-0.261***	-0.337***	-20.104***	-20.371***	-22.368***
	(0.045)	(0.042)	(0.037)	(0.027)	(0.028)	(0.034)	(0.863)	(0.692)	(0.473)
Disease Class FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Count of Publications	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	3385	4760	4756	3385	4760	4756	3385	4760	4756

Note: † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease-class level in parentheses.

Estimates from OLS models. The sample is at the disease-level. This table replicates our baseline specification using alternative windows to define the period of early search. We report the results including all years before 1995 (Columns (1), (4), and (7)), the baseline (Columns (2), (5), and (8)), and before 2005 (Columns (3), (6), and (9)). For each human disease, we determine the highest DisGeNET score for any gene identified during the exploration period. We classify maximum scores below the 60$^{\text{th}}$ percentile as L, scores between the 60$^{\text{th}}$ and 90$^{\text{th}}$ percentile as M, and scores above the 90$^{\text{th}}$ percentile as H. Columns 1-3 show the impact of early discoveries on the likelihood of finding any breakthrough during the sample period. Columns 4-6 show the impact on the number of new genes explored per scientific publication in the years following the exploration period. Columns 7-9 show the impact on the delay in discovering a breakthrough, defined as the years that elapsed from 1980 (the first year of our panel). In all models, diseases classified under L constitute the excluded category. We include disease-class fixed effects and control for the number of publications in the post-exploration period. See text for more details.

Table C.11: Using Share of Publications on Disease to Define the Exploration Period.

	Group Breakthrough			Group Exploration			Group Delay		
	(1) 5%	(2) 10%	(3) 15%	(4) 5%	(5) 10%	(6) 15%	(7) 5%	(8) 10%	(9) 15%
Max Found: M	-0.086***	-0.117***	-0.101**	-0.103**	-0.149***	-0.137***	3.619***	3.203***	2.645***
	(0.026)	(0.032)	(0.033)	(0.031)	(0.034)	(0.031)	(0.671)	(0.686)	(0.719)
Max Found: H	0.332***	0.456***	0.549***	-0.168***	-0.243***	-0.245***	-14.057***	-17.297***	-18.972***
	(0.035)	(0.040)	(0.037)	(0.022)	(0.025)	(0.025)	(0.937)	(0.893)	(0.842)
Disease Class FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Count of Publications	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	4761	4761	4761	4761	4761	4761	4761	4761	4761

Note: † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease-class level in parentheses.

Estimates from OLS models. The sample is at the disease-level. This table replicates our baseline specification using a disease-specific definition of the early exploration period. We compute the share of total publications on a given disease that were published by a specific year. We then define the end of the exploration period as the year 5% of publications were published (Columns (1), (4), and (7)), the year 10% of publications were published (Columns (2), (5), and (8)), and the year 15% of publications were published (Columns (3), (6), and (9)). For each human disease, we determine the highest DisGeNET score for any gene identified during the exploration period (which varies by disease). We classify maximum scores below the 60$^{\text{th}}$ percentile as L, scores between the 60$^{\text{th}}$ and 90$^{\text{th}}$ percentile as M, and scores above the 90$^{\text{th}}$ percentile as H. Columns 1-3 show the impact of early discoveries on the likelihood of finding any breakthrough during the sample period. Columns 4-6 show the impact on the number of new genes explored per scientific publication in the years following the exploration period. Columns 7-9 show the impact on the delay in discovering a breakthrough, defined as the years that elapsed from 1980 (the first year of our panel). In all models, diseases classified under L constitute the excluded category. We include disease-class fixed effects and control for the number of publications in the post-exploration period. See text for more details.

Table C.12: Using Alternative Windows to Examine Follow-on Exploration.

	New Genes/Papers			
	All Years	5 Years	10 Years	Until H
	(1)	(2)	(3)	(4)
Max Found: M	-0.144***	-0.246***	-0.204***	-0.172***
	(0.023)	(0.040)	(0.035)	(0.028)
Max Found: H	-0.261***	-0.454***	-0.353***	
	(0.028)	(0.046)	(0.031)	
Disease Class FE	Yes	Yes	Yes	Yes
Count of Publications	Yes	Yes	Yes	Yes
N	4760	4495	4715	1778

Note: † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease-class level in parentheses.

Estimates from OLS models. The sample is at the disease-level. This table replicates our baseline specification using alternative windows to evaluate exploration dynamics after a genetic discovery. We report the results from the baseline (Column (1)), the 5 subsequent years after the year 2000 (Column (2)), the 10 subsequent years after the year 2000 (Column (3)), and until the first high gene score is found (Column (4)). For each human disease, we determine the highest DisGeNET score for any gene identified during the exploration period (i.e., pre-2000). We classify maximum scores below the 60$^{\text{th}}$ percentile as L, scores between the 60$^{\text{th}}$ and 90$^{\text{th}}$ percentile as M, and scores above the 90$^{\text{th}}$ percentile as H. Each model shows the impact on the number of new genes explored per scientific publication in the years following the exploration period. In all models, diseases classified under L constitute the excluded category. We include disease-class fixed effects and control for the number of publications post-2000. See text for more details.

Table C.13: Considering only Diseases that Have a Breakthrough by the End of the Sample Period.

	Group Exploration		Group Delay
	(1) 5 Years After	(2) All Years After	(3) Years From 1980
Max Found: M	-0.142*	-0.074†	0.747
	(0.062)	(0.040)	(0.468)
Max Found: H	-0.460***	-0.220***	-13.588***
	(0.062)	(0.036)	(0.457)
Disease Class FE	Yes	Yes	Yes
Count of Publications	Yes	Yes	Yes
N	3442	3581	3581

Note: † $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Standard errors clustered at the disease-class level in parentheses.

Estimates from OLS models. The sample is at the disease-level. This table replicates our baseline specification removing any diseases without a breakthrough during the sample period. For each human disease, we determine the highest DisGeNET score for any gene identified during the exploration period (i.e., pre-2000). We classify maximum scores below the 60$^{\text{th}}$ percentile as L, scores between the 60$^{\text{th}}$ and 90$^{\text{th}}$ percentile as M, and scores above the 90$^{\text{th}}$ percentile as H. Column 1 shows the impact on the number of new genes explored per publication in the 5 years following the exploration window, while Column 2 shows the impact on the number of new genes explored per publication in all years following the exploration window. Column 3 shows the impact on the delay in discovering a breakthrough, defined as the years that elapsed from 1980 (the first year of our panel). In all models, diseases classified under L constitute the excluded category. We include disease-class fixed effects and control for the number of publications post-2000. See text for more details.

Table C.14: Predicting Scientific Value of Gene-Disease Pairs from Related Conditions

Panel A: Predicting H

	H Gene (0/1)
Max Sibling: L	0.011***
	(0.002)
Max Sibling: M	0.025***
	(0.002)
Max Sibling: H	0.231***
	(0.003)
N	810,377

Panel B: Predicting M

	M Gene (0/1)
Max Sibling: L	0.013***
	(0.002)
Max Sibling: M	0.144***
	(0.002)
Max Sibling: H	0.034***
	(0.003)
N	810,377

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Robust standard errors in parentheses.
This table examines whether a gene-disease pair is more likely to be classified as H (Panel A) or M (Panel B) based on the strength of the gene's associations with related diseases (sharing the same parent disease in the MeSH taxonomy). The data is at the gene-disease level. For each gene-disease pair, we record the highest DisGeNET score between the gene in question and any disease classified as a sibling of the focal disease. We classify maximum scores below the 60$^{\text{th}}$ percentile as L, scores between the 60$^{\text{th}}$ and 90$^{\text{th}}$ percentile as M, and scores above the 90$^{\text{th}}$ percentile as H. *H Gene*: 0/1=1 if the pair is classified as H, and *M Gene*: 0/1=1 if the pair is classified as M. In all models, gene-disease associations for which no sibling score is found constitute the excluded category.

D Experimental Instructions and Interfaces

D.1 No-Data Condition

Instructions

General Information

Welcome. This is an experiment in the economics of decision-making. If you pay close attention to these instructions, you can earn a significant amount of money paid to you at the end of the experiment via bank transfer.

To participate in this online experiment, you will need to use your notebook or personal computer (mobile phones are not supported). If you are using a device that is not supported, please copy the experiment link, open a notebook or pc and paste the link into the address bar.

Your computer screen will display useful information. Remember that the information on your computer screen is PRIVATE. To ensure best results for yourself and accurate data for the experimenters, please DO NOT COMMUNICATE or interact with other people on other media at any point during the experiment. If you have any questions, or need assistance of any kind, please call +43-678-780-7284 or use Zoom anytime during the experiment and one of the experimenters will help you privately. We expect the entire experiment to take up to 60 minutes to complete.

Following these instructions, you will be asked to make some choices. There are no correct choices. Your choices depend on your preferences and beliefs, so different participants will usually make different choices. You will be paid according to your choices, so read these instructions carefully and think before you decide.

The Basic Idea

There are 5 mountains and each of them hides one type of gem, which can only be found by exploring the mountain.

There are 3 types of gems hidden:

The exact values of the topazes, rubies, and diamonds vary across rounds but the diamonds are always worth more than the rubies and the rubies are always worth more than the topazes:

You choose which mountains to explore and the value of the gems you find are your earnings in dollars.

How the Gems Are Distributed

You will not know where the gems are hidden from the outset. At the beginning of every round, a gem for each mountain will be randomly drawn, so any gem could be hidden in any mountain.

For each mountain, there is a:

- 60% chance it contains a topaz 
- 20% chance it contains a ruby 
- 20% chance it contains a diamond 

These chances are the same for all five mountains. Hence, there is some chance that there could be more than one diamond, but there is also some chance that there could be no diamond. Further, even if, for example, the first two mountains happen to contain a diamond, the chance that the third mountain contains a diamond is still 20%.

How Participants Choose Mountains

In each round, participants choose which mountain to explore. The choice does not happen simultaneously, but participants choose sequentially, one after the other, according to a random order that changes every round. You can choose to explore any mountain you wish. If you choose the same mountain chosen by other participants, each of you will receive the gem's value uncovered. Similarly, if someone else chooses the same mountain that you previously chose, you will still receive the full gem's value (and so will the other participant(s) that chose it).
To repeat, no participant has any initial information in Stage 1 on the location of gems.

Each Round Has 2 Stages

A round consists of 2 stages. At the beginning of a new round, gems are redrawn for each of the five mountains. The position of gems will **not** be reset between the two stages in a round.
In Stage 1, all participants sequentially choose one mountain to explore. Before choosing a mountain, you will see which mountains have been selected by the other participants in your group who chose before you, and how many participants have selected each mountain. You can choose the same mountain or a different mountain.
At the end of Stage 1, the gems hidden in each mountain selected by all participants in Stage 1 are revealed, and you earn the value of the gem hidden in the mountain you chose.
In Stage 2, you can again choose any of the same five mountains; that is you can either choose the same mountain of Stage 1 or switch to another one. The position of gems remains the same as in Stage 1, but this time you will also see the gems located in the mountains revealed in Stage 1 in addition to the mapped mountain.
At the end of Stage 2, the gems hidden in each mountain selected by all participants in Stage 2 are revealed, and you earn the value of the gem hidden in the mountain you chose in Stage 2. You will also see your total earnings for the round which equals the sum of the value of the gem you found in Stage 1 and the value of the gem you found in Stage 2.

Game Structure

The game is divided into 4 blocks, each made of 5 rounds, with each round encompassing the two stages described above. At the beginning of each block, you will be randomly assigned to a new group of 5 participants, with whom you will play for the entire block (5 rounds in total). After the block is complete, you will be randomly assigned to a new group of 5 participants. Again, you will play for 5 rounds. This procedure will be repeated 4 times in total.
You will be reminded of this information in the top-right corner of your screen, as in the example below:

Payment

Fixed Participation Fee: You will earn a participation fee of $5.00 for participating in this experiment.

Additional Payment and Random Round: One round will be randomly selected for payment at the end of the experiment. You will be paid and your earnings in that round as described above. Any of the 20 rounds (4 blocks with 5 rounds each) could be the one selected, so you should treat each round as if it will be the one determining your payment.
This protocol of determining payments suggests that you should choose in each round as if it is the only round that determines your payment as the dollar value of the gems you select will directly translate into your earnings.

Survey and Payment: In addition to the participation fee and the payment for the randomly selected round, you will perform a small task at the very end of the experiment, and your earnings from this task will be paid to you.

You will be informed of your payment and the round chosen for payment at the end of the experiment. The $ you have earned will be converted into Euros at an exchange rate of $ 1 = € 0.67. Finally, after completing the experiment you will be paid electronically via bank transfer.

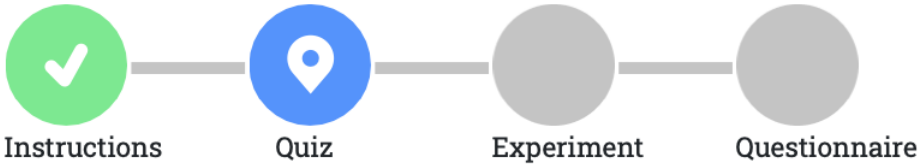
Frequently Asked Questions

Q1: Is this some kind of psychology experiment with an agenda you haven't told us?
A: No, it is an economics experiment. If we do anything deceptive or don't pay you as described, then you can complain to the University of Toronto Research Ethics Board and we will be in serious trouble. These instructions are meant to clarify how you earn money and our interest is in seeing how people make decisions.

Q2: Is there a "correct" or "wrong" choice of action? Is this kind of a test?
A: No, your optimal choice depends on your preferences and beliefs and different people may hold different beliefs.

Next

This button will be activated after 281 seconds. Please take your time to read through the instructions.



You have successfully finished reading the instructions.

The quiz, consisting of 8 questions in total, follows.

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q1: In each round, you will select two mountains (one in Stage 1, and one in Stage 2) and collect the gem that they hide. You can choose the same mountain in both stages, or change after Stage 1." :

- Correct
- Incorrect

Read Instructions

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q2: If more than one player selects the same mountain, they will all collect the full value of the gem." :

- Correct
- Incorrect

Read Instructions

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q3: At the beginning of a new round, the gems are redrawn for each mountain." :

- Correct
- Incorrect

Read Instructions

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q4: No group member has any private initial information in Stage 1 on the location of gems." :

- Correct
- Incorrect

Read Instructions

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q5: The position of gems will not be reset between the two stages of a round." :

- Correct
- Incorrect

Read Instructions

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q6: All group members select the mountains simultaneously." :

- Correct
- Incorrect

Read Instructions

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q7: If another group member chose a mountain before you, you cannot choose it again." :

- Correct
- Incorrect

Read Instructions

Next

Quiz Time!

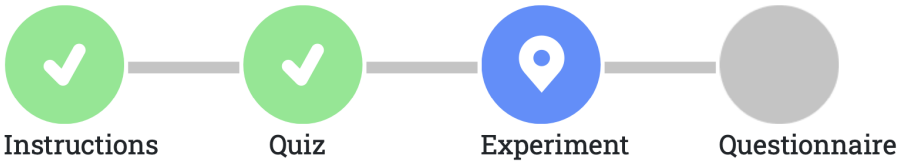
Please mark the following statements as correct/incorrect:

"Q8: At the end of the experiment, one round will be randomly selected for payment." :

- ○ Correct
- ○ Incorrect

Read Instructions

Next



You have successfully finished the quiz.

The experiment follows: When you are ready please click "Next" to start the experiment.

Next

Start of Block 1

This is Block 1 of 4 and each Block consists of 5 Rounds.

You have been randomly assigned to a **new** group of 5 participants.

Next

Start of Round 1

You are now in Round 1 of 5 and each Round consists of 2 Stages.

The computer redrew the gems for each mountain.

No participant has any initial information on the location of gems.

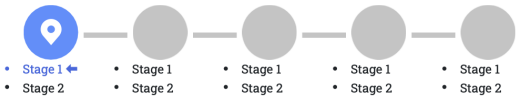
In this round, for each mountain, there could be:

: a topaz worth **$1.00** with **60%** chance
: a ruby worth **$6.00** with **20%** chance
: a diamond worth **$11.00** with **20%** chance

Next

Stage 1

This is Block 1 of 4: You are in Round 1 of 5.



In this round, for each mountain, there could be:

: a topaz worth **$1.00** with **60%** chance
: a ruby worth **$6.00** with **20%** chance
: a diamond worth **$11.00** with **20%** chance

The location of gems is random and no participant has any initial information where each gem is hidden.

It is NOT your turn yet, please wait.

Read Instructions

Stage 1

This is Block 1 of 4: You are in Round 1 of 5.

In this round, for each mountain, there could be:

: a topaz worth **$1.00** with **60%** chance
: a ruby worth **$6.00** with **20%** chance
: a diamond worth **$11.00** with **20%** chance

The location of gems is random and no participant has any initial information where each gem is hidden.

Now it is YOUR TURN, please select a mountain.

Read Instructions

Confirm your mountain choice

Stage 1: Earnings

You selected Mountain 3 and found a . Thus, you earned **$11.00** from your choice.

All discovered gems and their locations are highlighted below.
These will also be displayed in Stage 2 when you make your next choice.

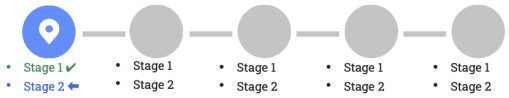
Click "Next" to proceed to the next stage.

Read Instructions

Next

Stage 2

This is Block 1 of 4: You are in Round 1 of 5.



In this round, for each mountain, there could be:

: a topaz worth **$1.00** with **60%** chance
: a ruby worth **$6.00** with **20%** chance
: a diamond worth **$11.00** with **20%** chance

Now it is YOUR TURN, please select a mountain.

Read Instructions

Confirm your mountain choice

Stage 2: Earnings

You selected Mountain 3 and found a . Thus, you earned **$11.00** from your choice.

Your total earnings from both stages in this round are **$11.00** + **$11.00** = **$22.00**

All discovered gems and their locations in both Stages are highlighted below.

Please click "Next" to proceed to the next round.

Read Instructions

Next

D.2 Data Condition

Instructions

General Information

Welcome. This is an experiment in the economics of decision-making. If you pay close attention to these instructions, you can earn a significant amount of money paid to you at the end of the experiment via bank transfer.

To participate in this online experiment, you will need to use your notebook or personal computer (mobile phones are not supported). If you are using a device that is not supported, please copy the experiment link, open a notebook or pc and paste the link into the address bar.

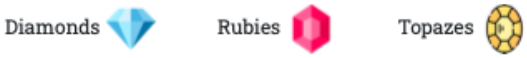
Your computer screen will display useful information. Remember that the information on your computer screen is PRIVATE. To ensure best results for yourself and accurate data for the experimenters, please DO NOT COMMUNICATE or interact with other people on other media at any point during the experiment. If you have any questions, or need assistance of any kind, please call +43-678-780-7284 or use Zoom anytime during the experiment and one of the experimenters will help you privately. We expect the entire experiment to take up to 60 minutes to complete.

Following these instructions, you will be asked to make some choices. There are no correct choices. Your choices depend on your preferences and beliefs, so different participants will usually make different choices. You will be paid according to your choices, so read these instructions carefully and think before you decide.

The Basic Idea

There are 5 mountains and each of them hides one type of gem, which can only be found by exploring the mountain.

There are 3 types of gems hidden:



The exact values of the topazes, rubies, and diamonds vary across rounds but the diamonds are always worth more than the rubies and the rubies are always worth more than the topazes:

You choose which mountains to explore and the value of the gems you find are your earnings in dollars.

How the Gems Are Distributed

You will not know where the gems are hidden from the outset. At the beginning of every round, a gem for each mountain will be randomly drawn, so any gem could be hidden in any mountain.

For each mountain, there is a:

- 60% chance it contains a topaz 
- 20% chance it contains a ruby 
- 20% chance it contains a diamond 

These chances are the same for all five mountains. Hence, there is some chance that there could be more than one diamond, but there is also some chance that there could be no diamond. Further, even if, for example, the first two mountains happen to contain a diamond, the chance that the third mountain contains a diamond is still 20%.

The Map

At the beginning of each round, **one** mountain will be randomly selected to be mapped and its gem will be uncovered to all participants. Each participant will be able to see the same gem contained by the mountain. The mountain chosen for mapping is random and changes in each round. Besides the map, no participant has any other initial information in Stage 1 on the location of gems.

How Participants Choose Mountains

In each round, participants choose which mountain to explore. The choice does not happen simultaneously, but participants choose sequentially, one after the other, according to a random order that changes every round. You can choose to explore any mountain you wish or select the mapped mountain. If you choose the same mountain chosen by other participants, each of you will receive the gem's value uncovered. Similarly, if someone else chooses the same mountain that you previously chose, you will still receive the full gem's value (and so will the other participant(s) that chose it).
To repeat, all participant have the same information in Stage 1 on the location of one of the gems.

Each Round Has 2 Stages

A round consists of 2 stages. At the beginning of a new round, gems are redrawn for each of the five mountains. The position of gems will **not** be reset between the two stages in a round.
Then, one of the mountains randomly selected for mapping and the gem it hides is revealed to all players.
In Stage 1, all participants sequentially choose one mountain to explore. Before choosing a mountain, you will see which mountains have been selected by the other participants in your group who chose before you, and how many participants have selected each mountain. You can choose the same mountain or a different mountain.
At the end of Stage 1, the gems hidden in each mountain selected by all participants in Stage 1 are revealed, and you earn the value of the gem hidden in the mountain you chose.
In Stage 2, you can again choose any of the same five mountains; that is you can either choose the same mountain of Stage 1 or switch to another one. The position of gems remains the same as in Stage 1, but this time you will also see the gems located in the mountains revealed in Stage 1 in addition to the mapped mountain.
At the end of Stage 2, the gems hidden in each mountain selected by all participants in Stage 2 are revealed, and you earn the value of the gem hidden in the mountain you chose in Stage 2. You will also see your total earnings for the round which equals the sum of the value of the gem you found in Stage 1 and the value of the gem you found in Stage 2.

Game Structure

The game is divided into 4 blocks, each made of 5 rounds, with each round encompassing the two stages described above. At the beginning of each block, you will be randomly assigned to a new group of 5 participants, with whom you will play for the entire block (5 rounds in total). After the block is complete, you will be randomly assigned to a new group of 5 participants. Again, you will play for 5 rounds. This procedure will be repeated 4 times in total.
You will be reminded of this information in the top-right corner of your screen, as in the example below:

Payment

Fixed Participation Fee: You will earn a participation fee of $5.00 for participating in this experiment.

Additional Payment and Random Round: One round will be randomly selected for payment at the end of the experiment. You will be paid and your earnings in that round as described above. Any of the 20 rounds (4 blocks with 5 rounds each) could be the one selected, so you should treat each round as if it will be the one determining your payment.
This protocol of determining payments suggests that you should choose in each round as if it is the only round that determines your payment as the dollar value of the gems you select will directly translate into your earnings.

Survey and Payment: In addition to the participation fee and the payment for the randomly selected round, you will perform a small task at the very end of the experiment, and your earnings from this task will be paid to you.

You will be informed of your payment and the round chosen for payment at the end of the experiment. The $ you have earned will be converted into Euros at an exchange rate of $ 1 = € 0.67. Finally, after completing the experiment you will be paid electronically via bank transfer.

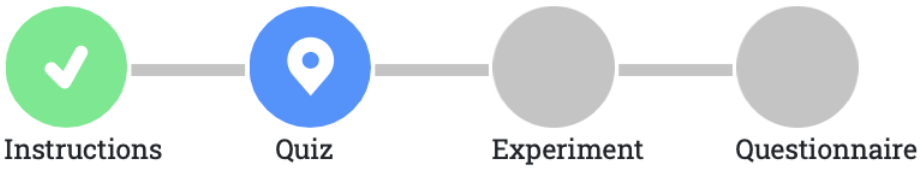
Frequently Asked Questions

Q1: Is this some kind of psychology experiment with an agenda you haven't told us?
A: No, it is an economics experiment. If we do anything deceptive or don't pay you as described, then you can complain to the University of Toronto Research Ethics Board and we will be in serious trouble. These instructions are meant to clarify how you earn money and our interest is in seeing how people make decisions.

Q2: Is there a "correct" or "wrong" choice of action? Is this kind of a test?
A: No, your optimal choice depends on your preferences and beliefs and different people may hold different beliefs.

Next

This button will be activated after 281 seconds. Please take your time to read through the instructions.



You have successfully finished reading the instructions.

The quiz, consisting of 8 questions in total, follows.

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q1: In each round, you will select two mountains (one in Stage 1, and one in Stage 2) and collect the gem that they hide. You can choose the same mountain in both stages, or change after Stage 1." :

- Correct
- Incorrect

Read Instructions

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q2: If more than one player selects the same mountain, they will all collect the full value of the gem." :

- Correct
- Incorrect

Read Instructions

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q3: At the beginning of a new round, the gems are redrawn for each mountain." :

- Correct
- Incorrect

Read Instructions

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q4: No group member has any private initial information in Stage 1 on the location of gems." :

- Correct
- Incorrect

Read Instructions

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q5: The position of gems will not be reset between the two stages of a round." :

- Correct
- Incorrect

Read Instructions

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q6: All group members select the mountains simultaneously." :

- Correct
- Incorrect

Read Instructions

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q7: If another group member chose a mountain before you, you cannot choose it again." :

- Correct
- Incorrect

Read Instructions

Next

Quiz Time!

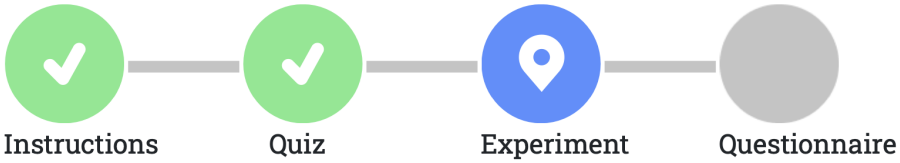
Please mark the following statements as correct/incorrect:

"Q8: At the end of the experiment, one round will be randomly selected for payment." :

- Correct
- Incorrect

Read Instructions

Next



You have successfully finished the quiz.

The experiment follows: When you are ready please click "Next" to start the experiment.

Next

Start of Block 1

This is Block 1 of 4 and each Block consists of 5 Rounds.

You have been randomly assigned to a **new** group of 5 participants.

Next

Start of Round 1

You are now in Round 1 of 5 and each Round consists of 2 Stages.

The computer redrew the gems for each mountain.

No participant has any initial information on the location of gems.

In this round, for each mountain, there could be:

: a topaz worth **$1.00** with **60%** chance
: a ruby worth **$6.00** with **20%** chance
: a diamond worth **$11.00** with **20%** chance

Next

Stage 1

In this round, for each mountain, there could be:

: a topaz worth **$1.00** with **60%** chance
: a ruby worth **$6.00** with **20%** chance
: a diamond worth **$11.00** with **20%** chance

The location of gems is random and no participant has any initial information where each gem is hidden.

It is NOT your turn yet, please wait.

Read Instructions

Stage 1

In this round, for each mountain, there could be:

: a topaz worth **$1.00** with **60%** chance
: a ruby worth **$6.00** with **20%** chance
: a diamond worth **$11.00** with **20%** chance

The location of gems is random and no participant has any initial information where each gem is hidden.

Now it is YOUR TURN, please select a mountain.

Read Instructions

Confirm your mountain choice

Stage 1: Earnings

You selected Mountain 2 and found a . Thus, you earned **$6.00** from your choice.

All discovered gems in addition to the mapped mountain and their locations are highlighted below.
These will also be displayed in Stage 2 when you make your next choice.

Click "Next" to proceed to the next stage.

Read Instructions

Next

Stage 2

In this round, for each mountain, there could be:

: a topaz worth **$1.00** with **60%** chance
: a ruby worth **$6.00** with **20%** chance
: a diamond worth **$11.00** with **20%** chance

Now it is YOUR TURN, please select a mountain.

Read Instructions

Confirm your mountain choice

Stage 2: Earnings

You selected Mountain 2 and found a  . Thus, you earned **$6.00** from your choice.

Your total earnings from both stages in this round are **$6.00** + **$6.00** = **$12.00**

All discovered gems and their locations in both Stages are highlighted below.

Please click "Next" to proceed to the next round.

Read Instructions

Next

D.3 Data Condition with Intermediate Rivalry

Instructions

General Information

Welcome. This is an experiment in the economics of decision-making. If you pay close attention to these instructions, you can earn a significant amount of money paid to you at the end of the experiment via bank transfer.

To participate in this online experiment, you will need to use your notebook or personal computer (mobile phones are not supported). If you are using a device that is not supported, please copy the experiment link, open a notebook or pc and paste the link into the address bar.

Your computer screen will display useful information. Remember that the information on your computer screen is PRIVATE. To ensure best results for yourself and accurate data for the experimenters, please DO NOT COMMUNICATE or interact with other people on other media at any point during the experiment. If you have any questions, or need assistance of any kind, please call +43-678-780-7284 or use Zoom anytime during the experiment and one of the experimenters will help you privately. We expect the entire experiment to take up to 60 minutes to complete.

Following these instructions, you will be asked to make some choices. There are no correct choices. Your choices depend on your preferences and beliefs, so different participants will usually make different choices. You will be paid according to your choices, so read these instructions carefully and think before you decide.

The Basic Idea

There are 5 mountains and each of them hides one type of gem, which can only be found by exploring the mountain.

There are 3 types of gems hidden:

The exact values of the topazes, rubies, and diamonds vary across rounds but the diamonds are always worth more than the rubies and the rubies are always worth more than the topazes:

You choose which mountains to explore and the value of the gems you find are your earnings in dollars.

How the Gems Are Distributed

You will not know where the gems are hidden from the outset. At the beginning of every round, a gem for each mountain will be randomly drawn, so any gem could be hidden in any mountain.

For each mountain, there is a:

- 60% chance it contains a topaz 
- 20% chance it contains a ruby 
- 20% chance it contains a diamond 

These chances are the same for all five mountains. Hence, there is some chance that there could be more than one diamond, but there is also some chance that there could be no diamond. Further, even if, for example, the first two mountains happen to contain a diamond, the chance that the third mountain contains a diamond is still 20%.

The Map

At the beginning of each round, **one** mountain will be randomly selected to be mapped and its gem will be uncovered to all participants. Each participant will be able to see the same gem contained by the mountain. The mountain chosen for mapping is random and changes in each round. Besides the map, no participant has any other initial information in Stage 1 on the location of gems.

How Participants Choose Mountains

In each round, participants choose which mountain to explore. The choice does not happen simultaneously, but participants choose sequentially, one after the other, according to a random order that changes every round. You can choose to explore any mountain you wish or select the mapped mountain. If you choose the same mountain already chosen by three other participants, you will **not** receive the gem's value uncovered. Instead you will receive a value of zero. Similarly, if someone else chooses the same mountain that you previously chose and you were among the first three to do so, you will receive the full gem's value (and the other participant(s) that chose it will not receive the gem's value uncovered if they were not among the first three participants to select that mountain).
To repeat, all participant have the same information in Stage 1 on the location of one of the gems.

Each Round Has 2 Stages

A round consists of 2 stages. At the beginning of a new round, gems are redrawn for each of the five mountains. The position of gems will **not** be reset between the two stages in a round.
In Stage 1, all participants sequentially choose one mountain to explore. Before choosing a mountain, you will see which mountains have been selected by the other participants in your group who chose before you, and how many participants have selected each mountain. You can choose the same mountain or a different mountain.
At the end of Stage 1, the gems hidden in each mountain selected by all participants in Stage 1 are revealed, and you earn the value of the gem hidden in the mountain you chose.
In Stage 2, you can again choose any of the same five mountains; that is you can either choose the same mountain of Stage 1 or switch to another one. The position of gems remains the same as in Stage 1, but this time you will also see the gems located in the mountains revealed in Stage 1 in addition to the mapped mountain.
At the end of Stage 2, the gems hidden in each mountain selected by all participants in Stage 2 are revealed, and you earn the value of the gem hidden in the mountain you chose in Stage 2. You will also see your total earnings for the round which equals the sum of the value of the gem you found in Stage 1 and the value of the gem you found in Stage 2.

Game Structure

The game is divided into 4 blocks, each made of 5 rounds, with each round encompassing the two stages described above. At the beginning of each block, you will be randomly assigned to a new group of 5 participants, with whom you will play for the entire block (5 rounds in total). After the block is complete, you will be randomly assigned to a new group of 5 participants. Again, you will play for 5 rounds. This procedure will be repeated 4 times in total.
You will be reminded of this information in the top-right corner of your screen, as in the example below:

Payment

Fixed Participation Fee: You will earn a participation fee of $5.00 for participating in this experiment.

Additional Payment and Random Round: One round will be randomly selected for payment at the end of the experiment. You will be paid and your earnings in that round as described above. Any of the 20 rounds (4 blocks with 5 rounds each) could be the one selected, so you should treat each round as if it will be the one determining your payment.
This protocol of determining payments suggests that you should choose in each round as if it is the only round that determines your payment as the dollar value of the gems you select will directly translate into your earnings.

Survey and Payment: In addition to the participation fee and the payment for the randomly selected round, you will perform a small task at the very end of the experiment, and your earnings from this task will be paid to you.

You will be informed of your payment and the round chosen for payment at the end of the experiment. The $ you have earned will be converted into Euros at an exchange rate of $ 1 = € 0.67. Finally, after completing the experiment you will be paid electronically via bank transfer.

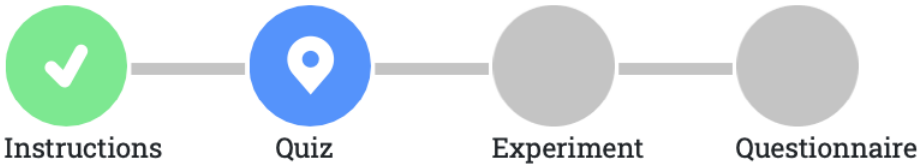
Frequently Asked Questions

Q1: Is this some kind of psychology experiment with an agenda you haven't told us?
A: No, it is an economics experiment. If we do anything deceptive or don't pay you as described, then you can complain to the University of Toronto Research Ethics Board and we will be in serious trouble. These instructions are meant to clarify how you earn money and our interest is in seeing how people make decisions.

Q2: Is there a "correct" or "wrong" choice of action? Is this kind of a test?
A: No, your optimal choice depends on your preferences and beliefs and different people may hold different beliefs.

Next

This button will be activated after 281 seconds. Please take your time to read through the instructions.



You have successfully finished reading the instructions.

The quiz, consisting of 8 questions in total, follows.

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q1: In each round, you will select two mountains (one in Stage 1, and one in Stage 2) and collect the gem that they hide. You can choose the same mountain in both stages, or change after Stage 1." :

- ○ Correct
- ○ Incorrect

Read Instructions

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q2: If more than one player selects the same mountain, all players will always collect the full value of the gem." :

- Correct
- Incorrect

Read Instructions

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q3: At the beginning of a new round, the gems are redrawn for each mountain." :

- Correct
- Incorrect

Read Instructions

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q4: No group member has any private initial information in Stage 1 on the location of gems." :

- Correct
- Incorrect

Read Instructions

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q5: The position of gems will not be reset between the two stages of a round." :

- Correct
- Incorrect

Read Instructions

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q6: All group members select the mountains simultaneously." :

- Correct
- Incorrect

Read Instructions

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q7: If another group member chose a mountain before you, you cannot choose it again." :

- Correct
- Incorrect

Read Instructions

Next

Quiz Time!

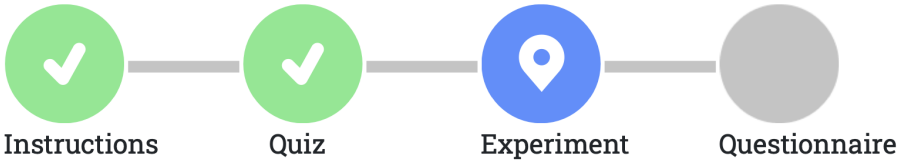
Please mark the following statements as correct/incorrect:

"Q8: At the end of the experiment, one round will be randomly selected for payment." :

- Correct
- Incorrect

Read Instructions

Next



You have successfully finished the quiz.

The experiment follows: When you are ready please click "Next" to start the experiment.

Next

Start of Block 1

This is Block 1 of 4 and each Block consists of 5 Rounds.

You have been randomly assigned to a **new** group of 5 participants.

Next

Start of Round 1

You are now in Round 1 of 5 and each Round consists of 2 Stages.

The computer redrew the gems for each mountain.

No participant has any initial information on the location of gems.

In this round, for each mountain, there could be:

: a topaz worth **$1.00** with **60%** chance
: a ruby worth **$6.00** with **20%** chance
: a diamond worth **$11.00** with **20%** chance

Next

Stage 1

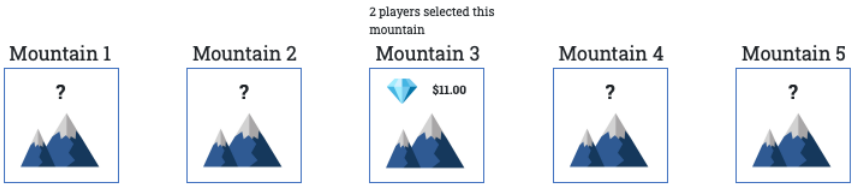
This is Block 1 of 4: You are in Round 1 of 5.

In this round, for each mountain, there could be:

: a topaz worth **$1.00** with **60%** chance
: a ruby worth **$6.00** with **20%** chance
: a diamond worth **$11.00** with **20%** chance

The location of gems is random and no participant has any initial information where each gem is hidden.

It is NOT your turn yet, please wait.



Read Instructions

Stage 1

This is Block 1 of 4: You are in Round 1 of 5.

In this round, for each mountain, there could be:

: a topaz worth **$1.00** with **60%** chance
: a ruby worth **$6.00** with **20%** chance
: a diamond worth **$11.00** with **20%** chance

The location of gems is random and no participant has any initial information where each gem is hidden.

Now it is YOUR TURN, please select a mountain.

Read Instructions

Confirm your mountain choice

Stage 1: Earnings

You selected Mountain 3 and found a 💎. Thus, you earned **$0.00** from your choice because you were **not among the first three players** to select this mountain.

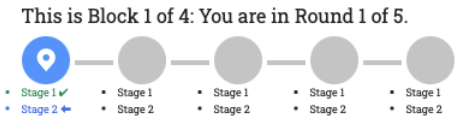
All discovered gems in addition to the mapped mountain and their locations are highlighted below.
These will also be displayed in Stage 2 when you make your next choice.

Click "Next" to proceed to the next stage.

Read Instructions

Next

Stage 2



In this round, for each mountain, there could be:

: a topaz worth **$1.00** with **60%** chance
: a ruby worth **$6.00** with **20%** chance
: a diamond worth **$11.00** with **20%** chance

Now it is YOUR TURN, please select a mountain.

Read Instructions

Confirm your mountain choice

Stage 2: Earnings

You selected Mountain 2 and found a  . Thus, you earned **$6.00** from your choice.

Your total earnings from both stages in this round are **$0.00** + **$6.00** = **$6.00**

All discovered gems and their locations in both Stages are highlighted below.

Please click "Next" to proceed to the next round.

Read Instructions

Next

D.4 No-Data Condition with Extreme Rivalry

Instructions

General Information

Welcome. This is an experiment in the economics of decision-making. If you pay close attention to these instructions, you can earn a significant amount of money paid to you at the end of the experiment via bank transfer.

To participate in this online experiment, you will need to use your notebook or personal computer (mobile phones are not supported). If you are using a device that is not supported, please copy the experiment link, open a notebook or pc and paste the link into the address bar.

Your computer screen will display useful information. Remember that the information on your computer screen is PRIVATE. To ensure best results for yourself and accurate data for the experimenters, please DO NOT COMMUNICATE or interact with other people on other media at any point during the experiment. If you have any questions, or need assistance of any kind, please call +43-678-780-7284 or use Zoom anytime during the experiment and one of the experimenters will help you privately. We expect the entire experiment to take up to 60 minutes to complete.

Following these instructions, you will be asked to make some choices. There are no correct choices. Your choices depend on your preferences and beliefs, so different participants will usually make different choices. You will be paid according to your choices, so read these instructions carefully and think before you decide.

The Basic Idea

There are 5 mountains and each of them hides one type of gem, which can only be found by exploring the mountain.

There are 3 types of gems hidden:

The exact values of the topazes, rubies, and diamonds vary across rounds but the diamonds are always worth more than the rubies and the rubies are always worth more than the topazes:

You choose which mountains to explore and the value of the gems you find are your earnings in dollars.

How the Gems Are Distributed

You will not know where the gems are hidden from the outset. At the beginning of every round, a gem for each mountain will be randomly drawn, so any gem could be hidden in any mountain.

For each mountain, there is a:

- 60% chance it contains a topaz 
- 20% chance it contains a ruby 
- 20% chance it contains a diamond 

These chances are the same for all five mountains. Hence, there is some chance that there could be more than one diamond, but there is also some chance that there could be no diamond. Further, even if, for example, the first two mountains happen to contain a diamond, the chance that the third mountain contains a diamond is still 20%.

How Participants Choose Mountains

In each round, participants choose which mountain to explore. The choice does not happen simultaneously, but participants choose sequentially, one after the other, according to a random order that changes every round. You can choose to explore any mountain you wish or select the mapped mountain. If you choose the same mountain already chosen by other participants, you will **not** receive the gem's value uncovered. Instead you will receive a value of zero. Similarly, if someone else chooses the same mountain that you previously chose and you were the first to do so, you will receive the full gem's value (and the other participant(s) that chose it will not receive the gem's value uncovered).
To repeat, no participant has any initial information in Stage 1 on the location of gems.

Each Round Has 2 Stages

A round consists of 2 stages. At the beginning of a new round, gems are redrawn for each of the five mountains. The position of gems will **not** be reset between the two stages in a round.
In Stage 1, all participants sequentially choose one mountain to explore. Before choosing a mountain, you will see which mountains have been selected by the other participants in your group who chose before you, and how many participants have selected each mountain. You can choose the same mountain or a different mountain.
At the end of Stage 1, the gems hidden in each mountain selected by all participants in Stage 1 are revealed, and you earn the value of the gem hidden in the mountain you chose.
In Stage 2, you can again choose any of the same five mountains; that is you can either choose the same mountain of Stage 1 or switch to another one. The position of gems remains the same as in Stage 1, but this time you will also see the gems located in the mountains revealed in Stage 1 in addition to the mapped mountain.
At the end of Stage 2, the gems hidden in each mountain selected by all participants in Stage 2 are revealed, and you earn the value of the gem hidden in the mountain you chose in Stage 2. You will also see your total earnings for the round which equals the sum of the value of the gem you found in Stage 1 and the value of the gem you found in Stage 2.

Game Structure

The game is divided into 4 blocks, each made of 5 rounds, with each round encompassing the two stages described above. At the beginning of each block, you will be randomly assigned to a new group of 5 participants, with whom you will play for the entire block (5 rounds in total). After the block is complete, you will be randomly assigned to a new group of 5 participants. Again, you will play for 5 rounds. This procedure will be repeated 4 times in total.
You will be reminded of this information in the top-right corner of your screen, as in the example below:

Payment

Fixed Participation Fee: You will earn a participation fee of $5.00 for participating in this experiment.

Additional Payment and Random Round: One round will be randomly selected for payment at the end of the experiment. You will be paid and your earnings in that round as described above. Any of the 20 rounds (4 blocks with 5 rounds each) could be the one selected, so you should treat each round as if it will be the one determining your payment.
This protocol of determining payments suggests that you should choose in each round as if it is the only round that determines your payment as the dollar value of the gems you select will directly translate into your earnings.

Survey and Payment: In addition to the participation fee and the payment for the randomly selected round, you will perform a small task at the very end of the experiment, and your earnings from this task will be paid to you.

You will be informed of your payment and the round chosen for payment at the end of the experiment. The $ you have earned will be converted into Euros at an exchange rate of $ 1 = € 0.67. Finally, after completing the experiment you will be paid electronically via bank transfer.

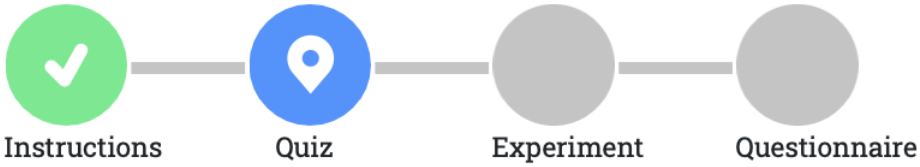
Frequently Asked Questions

Q1: Is this some kind of psychology experiment with an agenda you haven't told us?
A: No, it is an economics experiment. If we do anything deceptive or don't pay you as described, then you can complain to the University of Toronto Research Ethics Board and we will be in serious trouble. These instructions are meant to clarify how you earn money and our interest is in seeing how people make decisions.

Q2: Is there a "correct" or "wrong" choice of action? Is this kind of a test?
A: No, your optimal choice depends on your preferences and beliefs and different people may hold different beliefs.

Next

This button will be activated after 281 seconds. Please take your time to read through the instructions.



You have successfully finished reading the instructions.

The quiz, consisting of 8 questions in total, follows.

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q1: In each round, you will select two mountains (one in Stage 1, and one in Stage 2) and collect the gem that they hide. You can choose the same mountain in both stages, or change after Stage 1." :

- Correct
- Incorrect

Read Instructions

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q2: If more than one player selects the same mountain, all players will always collect the full value of the gem." :

- Correct
- Incorrect

Read Instructions

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q3: At the beginning of a new round, the gems are redrawn for each mountain." :

- Correct
- Incorrect

Read Instructions

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q4: No group member has any private initial information in Stage 1 on the location of gems." :

- Correct
- Incorrect

Read Instructions

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q5: The position of gems will not be reset between the two stages of a round." :

- Correct
- Incorrect

Read Instructions

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q6: All group members select the mountains simultaneously." :

- Correct
- Incorrect

Read Instructions

Next

Quiz Time!

Please mark the following statements as correct/incorrect:

"Q7: If another group member chose a mountain before you, you cannot choose it again." :

- Correct
- Incorrect

Read Instructions

Next

Quiz Time!

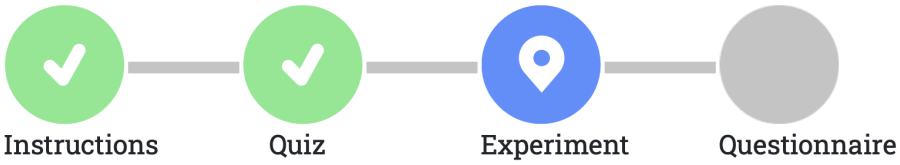
Please mark the following statements as correct/incorrect:

"Q8: At the end of the experiment, one round will be randomly selected for payment." :

- ○ Correct
- ○ Incorrect

Read Instructions

Next



You have successfully finished the quiz.

The experiment follows: When you are ready please click "Next" to start the experiment.

Next

Start of Block 1

This is Block 1 of 4 and each Block consists of 5 Rounds.

You have been randomly assigned to a **new** group of 5 participants.

Next

Start of Round 1

You are now in Round 1 of 5 and each Round consists of 2 Stages.

The computer redrew the gems for each mountain.

No participant has any initial information on the location of gems.

In this round, for each mountain, there could be:

: a topaz worth **$1.00** with **60%** chance
: a ruby worth **$6.00** with **20%** chance
: a diamond worth **$11.00** with **20%** chance

Next

Stage 1

This is Block 1 of 4: You are in Round 1 of 5.

In this round, for each mountain, there could be:

: a topaz worth **$1.00** with **60%** chance
: a ruby worth **$6.00** with **20%** chance
: a diamond worth **$11.00** with **20%** chance

The location of gems is random and no participant has any initial information where each gem is hidden.

It is NOT your turn yet, please wait.

Read Instructions

Stage 1

This is Block 1 of 4: You are in Round 1 of 5.

In this round, for each mountain, there could be:

: a topaz worth **$1.00** with **60%** chance
: a ruby worth **$6.00** with **20%** chance
: a diamond worth **$11.00** with **20%** chance

The location of gems is random and no participant has any initial information where each gem is hidden.

Now it is YOUR TURN, please select a mountain.

Read Instructions

Confirm your mountain choice

Stage 1: Earnings

You selected Mountain 3 and found a . Thus, you earned **$0.00** from your choice because you were **not the first player** to select this mountain.

All discovered gems in addition to the mapped mountain and their locations are highlighted below.
These will also be displayed in Stage 2 when you make your next choice.

Click "Next" to proceed to the next stage.

Read Instructions

Next

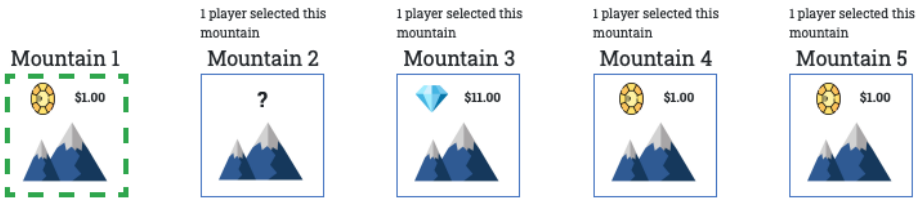
Stage 2

This is Block 1 of 4: You are in Round 1 of 5.

In this round, for each mountain, there could be:

: a topaz worth **$1.00** with **60%** chance
: a ruby worth **$6.00** with **20%** chance
: a diamond worth **$11.00** with **20%** chance

Now it is YOUR TURN, please select a mountain.



Read Instructions

Confirm your mountain choice

Stage 2: Earnings

You selected Mountain 1 and found a . Thus, you earned **$1.00** from your choice.

Your total earnings from both stages in this round are **$0.00** + **$1.00** = **$1.00**

All discovered gems and their locations in both Stages are highlighted below.

Please click "Next" to proceed to the next round.

Read Instructions

Next

D.5 Questionnaire and Risk-Preferences Elicitation Task

You have successfully finished the main part of the experiment.

A brief questionnaire together with a short task follows: When you are ready please click "Next".

Next

Please answer the following questions

Your answers will be kept confidential and will not affect your earnings for today's experiment.

Please state your age:

Please state your gender:

Please state your student type:

Please state your country of origin:

Please state your degree and field of study:

Please briefly explain, in your own words, the rules of today's experiment:

Please briefly describe how you reached your decisions in this experiment:

Please briefly describe how, in your opinion, other participants reached their decisions in this experiment:

Next

Instructions

Thank you for your participation so far. In the last task of the experiment, you will earn an additional reward based on a set of 10 choice problems.

How does it work?

The Choice: You will be asked to choose between two options, **"Option A"** and **"Option B"** where:

- "Option A" always pays `$4.00` with probability `p` and `$3.20` otherwise.
- "Option B" always pays `$7.70` with probability `p` and `$0.20` otherwise.

Repeated Choices:

- You will be asked to make a choice between "Option A" and "Option B" not once, but **ten** times where `p` will increase from `10%` to `100%`, `10%` at a time.

  For example, the first choice will have `p=10%` and you will choose whether you prefer "Option A" (`$4.00` with a `10%` chance or `$3.20` otherwise) or "Option B" (`$7.70` with a `10%` chance or `$0.20` otherwise).

- Each successive choice will increase `p` by `10` percentage points until the last choice where "Option A" will pay `$4.00` with certainty, and "Option B" will pay `$7.70` with certainty.

Note: Once you switch from choosing "Option A" to "Option B", it makes sense that you will continue to choose "Option B" in all consecutive choice problems. For example, if you prefer "Option B" when `p=80%`, then it makes sense to prefer "Option B" when `p=90%` and when `p=100%`, since "Option B" is even more attractive in these choice problems.

Therefore, we have designed the interface so that you must either (a) **always** choose "Option A" or "Option B" for all 10 choice problems or (b) if you **switch** to "Option B" for a given probability `p`, then you must choose "Option B" for all the following choices as well.

You can adjust your choices as many times as you wish. When you are ready to submit your choices, you can click on the "Next" button at the bottom of the page.

Payment

The computer will randomly select one of the 10 choice problems and pay you according to your choice in that problem where the computer will decide the outcome based on the value of `p`.

Next

Please Choose Between "Option A" and "Option B" on Every Line

	Option A		Option B	
	**$4.00** with a chance of 10%, **$3.20** otherwise	○ ○	**$7.70** with a chance of 10%, **$0.20** otherwise	
	**$4.00** with a chance of 20%, **$3.20** otherwise	○ ○	**$7.70** with a chance of 20%, **$0.20** otherwise	
	**$4.00** with a chance of 30%, **$3.20** otherwise	○ ○	**$7.70** with a chance of 30%, **$0.20** otherwise	
	**$4.00** with a chance of 40%, **$3.20** otherwise	○ ○	**$7.70** with a chance of 40%, **$0.20** otherwise	
	**$4.00** with a chance of 50%, **$3.20** otherwise	○ ○	**$7.70** with a chance of 50%, **$0.20** otherwise	
	**$4.00** with a chance of 60%, **$3.20** otherwise	○ ○	**$7.70** with a chance of 60%, **$0.20** otherwise	
	**$4.00** with a chance of 70%, **$3.20** otherwise	○ ○	**$7.70** with a chance of 70%, **$0.20** otherwise	
	**$4.00** with a chance of 80%, **$3.20** otherwise	○ ○	**$7.70** with a chance of 80%, **$0.20** otherwise	
	**$4.00** with a chance of 90%, **$3.20** otherwise	○ ○	**$7.70** with a chance of 90%, **$0.20** otherwise	
	$4.00 with a chance of 100%, **$3.20** otherwise	○ ○	**$7.70** with a chance of 100%, **$0.20** otherwise	

Read Instructions

Next

D.6 Payment Information

Thank you for participating in this experiment!

Your payoffs for this experiment are as follows:

Main Experiment:

- **Round 1** of **Block 1** was randomly selected for payment.
- In Stage 1, you found a 💎 and received **$11.00** and in Stage 2, you found a 💎 and received **$11.00**
- Thus, your total payoff is $11.00 + $11.00 = **$22.00**

Last Task of Experiment:

- The following choice problem was randomly selected:

	Option A		Option B
	$4.00 with a probability of 10%, $3.20 otherwise	◉ ○	$7.70 with a probability of 10%, $0.20 otherwise

- As indicated above, you chose **Option A**. The computer drew a random number to determine your payoff according to the chances specified.
- Your payoff is **$3.20**

Participation Fee:

- You earned a fee of **$5.00**

In total, you earned **$22.00 + $3.20 + $5.00 = $30.20** from your choices.

In euros, this corresponds to: **€20.13**

Please click Next to upload your bank details.

Next