Bread Upon the Waters: Corporate Science and the Benefits from Follow-On Public Research

Dror Shvadron*

September 20, 2023

Click here for the latest version.

Abstract

Why do firms produce scientific research and make it available to the public, including their rivals? An important but hitherto ignored benefit is that it can influence the direction of research conducted by external scientists in ways that benefit the focal firm. I show that external scientists often build upon a firm's publications, producing follow-on findings, which the firm then incorporates into its own future innovations. To account for the unobserved quality of the science involved, I develop a new instrumental variable that relies on the quasi-random assignment of accepted manuscripts to specific issues of scientific journals. Some publications attract more academic attention simply because they appear alongside contributions from prominent authors in the same journal issue. Using data on scientific publications by public firms between 1990 and 2012, I find that follow-on research not only drives firms' subsequent investments in science but also improves their patenting outcomes. The benefits are more pronounced for technological leaders, firms with complementary assets, and those operating in emerging research fields. In addition to being a valuable input into the firm's innovations, follow-on research also helps validate the quality of the firm's internal science, especially when there is greater uncertainty surrounding said science. My findings contribute to the understanding of why firms participate in public science.

*Duke University, Fuqua School of Business. E-mail: dror.shvadron@duke.edu.
I am particularly indebted to Ashish Arora, Wesley M. Cohen, Sharon Belenzon, and Bronwyn H. Hall for extensive advice and guidance. For thoughtful suggestions, I would like to thank Matt Marx, Ines Black, Sharique Hasan, Brian Silverman, Brent Goldfarb, Michaël Bikard, Daniel Souza (discussant), Arvids Zeidonis (discussant), Markus Perkmann (discussant), Keyvan Vakili (discussant), Annamaria Conti (discussant) and numerous participants at the Workshop on The Organization, Economics and Policy of Scientific Research 2023, the Open Innovation in Science Research Conference 2023, Strategy Science Conference 2023, and the CCC Doctoral Conference 2023. Lastly, I would like to thank my fellow colleagues at the Fuqua Strategy Area.

Cast your bread upon the waters,
for you will find it after many days.

Ecclesiastes 11

1 Introduction

Why do firms produce scientific research and make it available to the public, including their rivals? I find that firms' scientific publications are infrequently cited by their own patents. These publications do, however, influence external scientific research, which these firms later utilize. Prior literature highlighted the direct benefits of corporate science (Arora et al., 2021; Cohen & Levinthal, 1990; Fleming & Sorenson, 2004; Rosenberg, 1990). In this paper, I show that there are additional indirect benefits due to academics who build upon firms' publications. Such follow-on research drives the focal firms' subsequent investments in science and improves their patenting outcomes. Beyond providing inputs for subsequent innovation, follow-on research also validates the quality of the firms' own science, the latter being more important under greater uncertainty. Firms that are technological leaders or possess complementary assets are more likely to benefit. My findings suggest that by participating in public research, firms can strategically shape the knowledge environment in which they operate (Gavetti et al., 2017; Helfat et al., 2023). I contribute to understanding the determinants of firms' participation in public research and how the scientific community drives corporate innovation (Cohen et al., 2002; Jaffe, 1989; Mansfield, 1991).

Firms have various channels by which they can influence public research. Prior literature focused on direct ties, such as geographic proximity (Sohn, 2021), funding (Babina et al., 2023), and research collaborations (Bikard et al., 2019; Cockburn & Henderson, 1998). Firms can establish these ties by, for example, attending and sponsoring academic conferences (Baruffaldi & Poege, 2022). However, a primary channel of engaging with the scientific community is the disclosure of findings. By publishing scientific papers, firms can influence the direction of scientific research, even without establishing direct ties with academics. At the firm level, scientific publications correlate with higher market values (Simeth & Cincera, 2016), which is consistent with knowledge spillovers from corporate R&D having a positive effect on firms' performance (Alnuaimi & George, 2016; Belenzon, 2012; Bloom et al., 2013). The literature, however, lacks evidence regarding the mechanisms by which a firm's scientific publications can mobilize external resources to the firm's benefit (Alexy et al., 2013).

In this paper, I ask how firms benefit from external research that builds upon their own investments in science. I define participation in public research as the decision to invest in science and publicly disclose scientific findings. The prospect of valuable follow-on research from external sources can incentivize firms to participate in public research in the first place. I use the DISCERN database on scientific publications and patents of publicly listed U.S.-based firms between 1990 and 2012 (Arora, Belenzon, & Sheer, 2020), matched to data from Microsoft Academic Graph (MAG), Dimensions.ai, the American Men and Women of Science (AMWS) directory, and several complementary datasets.[1] I measure external follow-on research using scientific citations through multiple generations. Using these data, I test whether external follow-on research drives subsequent scientific publishing, scientist hiring, and patenting by the originating firms. Then, I explore the conditions that moderate these effects and the mechanisms that enable them.

An example illustrates the potential value of external follow-on findings. In 1986, two IBM researchers, Müller and Bednorz, made a breakthrough discovery. They were the first to find a material that behaves as a superconductor in high temperatures (above 77 Kelvin). A year later, they were awarded the Nobel Prize for their discovery. According to IBM's website:

> The scientific community shook. Scientists from across the world reproduced, modified and improved Müller and Bednorz's process at a breakneck pace—reigniting global interest in superconductors and accelerating superconductor development. Based on Müller and Bednorz's discovery, scientists soon developed materials that... opened the door for a multitude of practical applications.[2]

Indeed, in sixteen of its patents, IBM cited the original paper by Bednorz and Müller (1986). Meanwhile, up until 2015, the paper was cited by 6,166 scientific publications by authors unaffiliated with IBM. In turn, about 64,000 additional publications cited these publications, and 381,000

[1] I thank Hansen Zhang for sharing a match between AMWS and DISCERN data. I thank Bernardo Dionisi for sharing a match of DISCERN to Microsoft Academic Graph.

[2] IBM's 100 Icons of Progress. "High-Temperature Superconductors." Retrieved from https://www.ibm.com/ibm/history/ibm100/us/en/icons/hightempsuperconductors/.

publications cited the latter. On top of IBM's directly citing patents, 563 additional IBM patents cited these three generations of follow-on publications. Based on patent value estimates from Kogan et al. (2017), IBM's private value associated with the original patents was \$146 million. An additional \$1.9 billion is associated with IBM's patents related to the follow-on research. IBM's ability to capture value from external follow-on research contributes to a part of this latter figure.[3]

External follow-on research can be valuable for firms in various ways. First, it can serve as an input into firms' subsequent innovation. Inputs from the scientific community expand firms' scientific and inventive capabilities beyond the scope of the knowledge and expertise of their employees. In some cases, firms' publications of upstream science can influence academics to develop downstream applications (Ahmadpoor & Jones, 2017). In other cases, academics can broaden upstream basic science in response to downstream challenges that firms face. Under this view, firms' participation in public research is a way to expand the opportunities for the recombination of scientific knowledge (Fleming & Sorenson, 2004).

Second, firms can use follow-on research as a means to foster direct ties with academics, such as hiring, research collaborations, and funding. A vast literature explores the success factors of university-industry collaborations (Perkmann et al., 2021). However, any direct collaboration between a firm and external parties requires the firm to know in advance what expertise it requires. To that end, scientific publications can provide a method for a broad knowledge search (Leiponen & Helfat, 2010) that does not require prior acquaintance. A valuable follow-on finding by external actors can guide the firm toward direct channels of collaboration.

Third, external follow-on research can be valuable even when not used as an input. Due to the uncertainty that typically accompanies scientific research, scientists tend to rely on signals that point to promising research trajectories (Azoulay et al., 2014; Jin et al., 2021). Corporate R&D

[3] To this day, R&D managers at IBM recognize the value of engaging with scholars: *The key, emphasizes Jeff Welser, a VP and Lab Director at IBM's Almaden Research Center, is to be seen as an active participant and not just a spectator. "Being immersed is incredibly critical," he stresses. "You need to have people active at conferences, writing papers, and helping the field advance. You have to put value in to get value out." ... It's worth the effort. As IBM's Welser explains, "The thing that's nice about that kind of setup is that you get to pool your resources with government, academia, and other industry players, which is a good thing at the pre-competitive stage." Of course, once the collective breakthroughs become exciting new products, IBM and competitors like Intel and Microsoft fight like dogs.* Greg Satell, Innovative Companies Get Their Best Ideas from Academic Research — Here's How They Do It. (2016, April 21). Retrieved from https://hbr.org/2016/04/innovative-companies-get-their-best-ideas-from-academic-research-heres-how-they-do-it.

managers facing resource allocation decisions can use follow-on research as a signal for the quality of individuals' work and the promise of research trajectories within the firm. These signals can therefore redirect managerial attention and influence subsequent R&D investments by the firm. This argument is consistent with the view that external validation of research quality can reduce uncertainty around nascent research initiatives (Cockburn & Henderson, 1998).[4]

Descriptively, I show that the relation between corporate science and patenting is more substantial than that which is revealed by only considering patents that directly cite the firms' science. In my data, only 7% of corporate publications were cited internally by patents of the same firm. However, by observing external follow-on research, I find that an additional 33% of firms' publications were eventually indirectly cited by the same firms' patents. It is important to note that these citations require a long time horizon. On average, a third-generation follow-on paper is cited by a patent 13 years after the publication of the original paper. These findings suggest a close relation between firms' investments in science and their inventive activities and highlight the potential role of external research in supporting this relation. In addition, the findings suggest that external research might require years to evolve to the point where it is internally useful for the originating firm.

The challenge in identifying the effect of follow-on research on firms' outcomes is to control for the unobserved scientific quality of publications. Highly promising scientific findings would likely receive more interest from the scientific community, along with subsequent investments by the firm. To account for this source of endogeneity, I implement a new instrumental variable that exploits exogenous variation in the level of attention the scientific community pays a given publication. The instrument relies on the fact that in most scientific journals, the grouping of papers into journal issues follows a quasi-random "first-in-first-out" process. Within the same journal in a given year, some issues include publications by more prominent authors than other issues. At least until academic readership moved online, these issues were likely to draw more attention

[4]In a recent blog post, a past employee at Google Brain claimed that "Google's researcher promotion criteria were for some time linked to external recognition of research significance." (Lee, B. K. Why Did Google Brain Exist? April 2023. Retrieved from www.moderndescartes.com/essays/why_brain/.) The author further argued that, despite the inclusion of senior researchers on promotion committees, these panels have a limited capacity to accurately assess the scientific quality of research.

from fellow academics. The increased attention due to the inclusion of prominent authors drove citations to other papers in the same journal issue. To measure prominence, I calculate the H-index of all authors in journal issues in the sample. Next, I identify the top two prominent authors within each issue, excluding the authors of the focal papers. I use the sum of their H-indexes as an instrumental variable for follow-on research when estimating the effect on firms' scientific and innovation outcomes.

In the main analysis, I apply a two-stage least square estimation to identify the effect of follow-on research on firms' investments in science and patenting outcomes. First, I find that external follow-on research drives subsequent scientific investments by the originating firms. It increases scientific publishing, research collaborations, and academic conference participation by the corporate authors of the original publications. Second, follow-on research drives the firms to hire renowned scientists whose work is related to the focal publication. Third, follow-on research also improves the firms' patenting outcomes. It drives subsequent patenting by the authors and patent citations by the firm to the focal publication. Taken together, I interpret these results as evidence for the positive value of follow-on research for firms' scientific and innovative efforts. I complement these findings by reporting patent- and firm-level correlational evidence that follow-on research is related to firms' innovative and financial performance.

Next, I examine the moderators and mechanisms that drive the positive effects of follow-on research. I find that the role of follow-on research as quality validation is more beneficial for non-prominent focal authors, i.e., when quality uncertainty is high. Lastly, I explore the conditions that enable firms to benefit from follow-on research. I find more substantial effects in areas where firms face low competition, own intellectual property (IP) rights, and have internal research capabilities. Follow-on research is also more valuable in nascent scientific domains and areas where government funding is readily available.

I contribute to the literature on the determinants of firms' participation in public research. Investments in science enhance firms' combinative capabilities by extending their knowledge base and providing direct inputs into invention (Arora et al., 2021; Fleming & Sorenson, 2004; Kogut & Zander, 1992). In addition, they have a role in improving firms' absorptive capacity (Cohen &

Levinthal, 1990; Rosenberg, 1990). Both these views consider how firms are affected by *already-existing* external knowledge. However, given the magnitude of resources and outputs produced by the scientific community, it is evident that, on top of the benefits of better *access* to public research, firms can benefit from the ability to *influence* academics' research agenda. This paper highlights that, by participating in public research, firms can potentially influence the future content produced within the scientific information networks in ways that are privately beneficial.

My second contribution is to the literatures on open innovation and knowledge spillovers. Chesbrough (2003) argued that opening internal resources to external use and adopting external technologies can improve the performance of innovative firms. Later works, focusing on open-source software, showed that selective revealing could drive valuable external contributions (e.g., Alexy et al. (2013), Dahlander and Wallin (2006), and Henkel (2006)). Studies on knowledge spillovers from patent disclosures have shown the potential value of reabsorbing external developments (Alnuaimi & George, 2016; Belenzon, 2012; Yang et al., 2010; Yang & Steensma, 2014). However, while the disclosure of patented inventions is mandated by law, the decision to engage with the scientific community is a strategic choice (Alexy et al., 2018). This paper supports the view that opening up internal knowledge through participation in public research could be a strategic choice that drives down R&D costs, enhances the value of complementary assets, and allows firms to mitigate uncertainty associated with their investments in research.

2 Theoretical Framework

2.1 Firms' Participation in Public Research

Firms participate in public research by conducting scientific research and publicly disclosing their findings. Investments in science create knowledge that firms can use as an input into invention (Arora et al., 2021). Even when not used directly, scientific research enables firms to overcome the limitations of incremental search (Fleming & Sorenson, 2004) and improves their combinative capabilities (Arthur, 2011; Kogut & Zander, 1992). In addition, investments in science enhance firms' absorptive capacity, or their ability to identify, assimilate, and exploit *already-existing* external knowledge (Cohen & Levinthal, 1990; Rosenberg, 1990).

The publication of findings is typically an integral part of the scientific endeavor. Accordingly, it has long been recognized that, in some settings, firms encourage the disclosure of research findings (Cockburn & Henderson, 1998; Hicks, 1995). This occurs notwithstanding the potential risks of generating unintended knowledge spillovers and hindering firms' ability to capture the full returns from their R&D efforts (Arrow, 1962; Dasgupta & David, 1994; Nelson, 1959).[5] I argue that participation in public research, through investments and publication of research findings, is an important channel by which firms can influence the scientific community and eventually benefit from its resources.

2.2 Engaging with the Scientific Community

Firms' participation in public research can influence the pace and trajectory of scientific advances beyond their organizational boundaries. By participating, firms can potentially steer the focus of external scientific inquiry toward areas relevant to their needs. This capability has taken on heightened significance in recent times. The modern innovation ecosystem has experienced a shift toward a division of innovative labor among academia, incumbents, and startups (Arora & Gambardella, 1994). Accordingly, corporate innovation is increasingly reliant on public science (Fleming et al., 2019). The interconnected dynamics of innovative organizations are underpinned by firms' strategic influence on public research. Such influence can shape both immediate technological applications and the long-term scientific agenda.

Firms have various channels to engage with academics at universities and other public research institutions (Cohen et al., 2002). Many of these channels, and the focus of prior literature, require direct interpersonal ties between firms and academics. For example, university-industry research collaborations (UIC) enable firms to direct academic inquiry in their favor and benefit from academics' expertise and resources (Bikard et al., 2019).[6] Other direct channels, such as geo-

[5]Many have studied additional incentives to publish (Rotolo et al., 2022). The literature lists incentives such as employees' interest in publishing (Stern, 2004), the supportive role of publishing for IP strategy (Baker & Mezzetti, 2005), the effect on firms' reputation among regulators, investors, and suppliers (Baruffaldi et al., 2023; Harhoff et al., 2003; Polidoro & Theeke, 2012), and marketing strategies (Simeth & Raffo, 2013). These explanations do not directly relate scientific publishing to the accumulation of valuable knowledge by firms.

[6]See Perkmann et al. (2021), Perkmann et al. (2013) for recent literature reviews regarding UICs.

graphic proximity, funding, conference participation, and corporate hiring, are additional channels by which firms can influence academia (Babina et al., 2023; Baruffaldi & Poege, 2022; Sohn, 2021).

While direct ties with academics allow firms to influence their work, corporate publications are an indirect channel of engagement with the scientific community at large. They can lead to follow-on external research, defined as science outside the firm, informed by the findings they disclose. Given the cumulative nature of scientific research (Nelson & Winter, 1982), firms' publications can influence the research trajectory of academics outside the firm and lead them to invest their time and resources in follow-on work (Hicks, 1995).[7] In some cases, follow-on research creates downstream "scientific steps," cumulative findings that open up potential applications of the underlying science (Ahmadpoor & Jones, 2017; Dasgupta & David, 1994; David, 1998). In other cases, corporate science informs academics regarding downstream demand and technical problems that require upstream exploration. Follow-on research could emerge from distant geographic locations and a wide range of knowledge domains. In this sense, corporate publications as a channel of engagement differ from direct channels because they offer a broad, undirected search (Leiponen & Helfat, 2010).

2.3 Firms' Benefits from External Follow-On Research

Aside from the risks of knowledge spillovers to competitors, there are cases where knowledge spillovers from firms' R&D investments can also have positive effects on the originating firms. For example, knowledge can spill over to technologically-related firms not competing in the same product market (Bloom et al., 2013). Several recent studies explored firms' ability to benefit from spilled knowledge through reabsorption. Yang et al. (2010) used a sample of patents originating from telecommunications firms and showed a positive relationship between the spillover knowledge pool and firms' innovative activities. Belenzon (2012) studied how firms' ability to reabsorb knowledge spillovers from patents affects firm-level performance outcomes. Subsequently, Alnuaimi and George (2016) and Yang and Steensma (2014) studied how technology, firm, and industry characteristics interact with the ability to reabsorb knowledge. However, the positive effects

[7] Hicks (1995) referred to this as "collateral research."

described in this line of work might be partly unintentional, as patent law mandates the disclosure of information once an inventor seeks patent protection of intellectual property rights.

Participation in public research, as a case of selective knowledge revealing (Alexy et al., 2013), can be seen as a deliberate choice to engage in strategic openness (Alexy et al., 2018; Henkel, 2006). Openness, through the influence on external actors, can reduce R&D costs and increase value appropriation from complementary assets owned by the firm. Two recent works explore these arguments. Studying corporate patents in artificial intelligence (AI), Shen (2022) compares the cumulative innovation that follows patent-paper pairs to that of patents that lack a corresponding scientific publication. The results suggest that scientific publications broadcast the firms' inventions to a broader audience of external inventors. Also focusing on the AI industry, Jee and Sohn (2023) find correlational evidence that corporate publications influence the knowledge spillover pool and subsequently benefit firms' patenting outcomes.

2.3.1 What mechanisms make follow-on research beneficial to the originating firms?

Several mechanisms make external follow-on research beneficial for the originating firm. First, follow-on research can be useful as inputs into subsequent R&D activities. In many cases, academics have the expertise and resources that the originating firms lack. By inducing follow-on research, firms positioned to appropriate value from publicly-available knowledge can benefit from tapping into these resources. As a result of follow-on research, these firms will likely continue to invest in related research and experience better patenting outcomes.

Following scientific and patent citations within the data provides examples of these processes.[8] Universal Display Corporation (UDC) is a leading developer of organic LED (OLED) technologies. The company regularly publishes scientific papers. These disclosures enable others outside the firm to develop technologies related to OLED displays. In 2005, a paper by researchers at the Hong Kong University of Science and Technology suggested that a silver anode can improve the color saturation of OLED displays. Given their list of references, it is clear that UDC's prior publications informed the research group in Hong Kong. Following this publication, inventors within UDC

[8] See Appendix A for further details and additional examples from the data.

were able to further develop this idea. Eventually, in 2007, the firm patented a method to improve color saturation by adding a thin layer of silver to the design. Importantly, their patent cites the external paper. In this example, UDC's scientific publications led to privately-useful external follow-on research, even without direct ties between the firm and the external research group.

Second, direct research collaborations and the hiring of scientists require firms to know the identity of those with whom they wish to engage. As theorized by Alexy et al. (2013), follow-on research could provide firms with pathways toward valuable direct collaborations. Scientists that produce relevant follow-on research, or others that operate in closely-related fields, are good candidates for subsequent collaborations and recruitment. Therefore, follow-on research can lead to subsequent research collaborations, invention collaborations, and the hiring of academics by the originating firm.

AVI BioPharma[9], a biotech firm located in Cambridge, Massachusetts, developed a general-purpose method for altering gene expressions in RNA molecules. Following the publication of their method in 1997, a research group based at the University of Western Australia discovered that this method could potentially treat Duchenne, a type of muscular dystrophy. Their finding initiated further research on Duchenne treatments by the originating firm. It also spurred research collaborations between the firm and the Australian research group and gave rise to patent licensing agreements. Recently, the FDA approved the treatment, and it became commercially available for patients. In this case, the scientific community provided the required domain expertise that allowed the firm to find applications for their general-purpose technology. The follow-on findings oriented the firms' subsequent product development and resulted in direct university-industry collaborations.

Even if follow-on research is not used as input by the firm or does not lead to other forms of engagement, it could signal the technical promise and relevance of research trajectories and direct R&D managers' resource allocation decisions.[10] Scholarly attention, in the form of follow-

[9]Today named Sarepta Therapeutics, Inc.

[10]At the individual level, Azoulay et al. (2014) have documented a positive effect of scientists' status on the attention they receive from the scientific community. In recent work, Jin et al. (2021) have found that scientific prizes gave rise to extraordinary growth in research on related scientific topics, possibly because they signal to scientists areas of intellectually solid research domains and professional attractiveness. Plausibly, status plays a similar role within the firm. Prato and Ferraro (2018) found that firms' hiring of high-status individuals can disrupt

on citations, can act as a signal for the quality of an individual's work. Therefore, managers in charge of allocating resources for R&D can use scholarly citations as indicators of the quality of both individual researchers and their research paths. Such signals can guide their focus and affect future R&D investments. The value of follow-on research as a signal of quality is likely most important when there is more significant uncertainty regarding the quality of the work done by researchers employed by the firm.

2.3.2 Under what conditions is follow-on research beneficial to the originating firms?

Various factors likely affect firms' abilities to benefit from follow-on research. First, the competitive landscape in which the firm operates can affect the extent to which the firm can appropriate value from publicly-available knowledge. Given the public good nature of published science (Dasgupta & David, 1994), firms who face intense competition can less likely benefit from follow-on research. In contrast, firms that are technological leaders in a research domain and have stronger patenting experience compared to their competitors are likely better positioned to absorb and eventually benefit from external findings.

Second, when a firm owns complementary assets, such as patent protection over related intellectual property, it likely has an advantage over rivals in appropriating value from related follow-on research. Therefore, if the focal publication is part of a patent-paper pair (Marx & Fuegi, 2022), then follow-on research should have more significant effects on the firm's outcomes. In addition, a persistent internal research capability should improve firms' ability to absorb follow-on research when it becomes available. Outsourcing research to external academics, as is often the case in university-industry collaborations, is expected to hinder the firm's ability to benefit from follow-on research in subsequent years.

Third, characteristics of the scientific domain can also influence the benefits of follow-on research. In mature areas where many scholars are already operating, an additional publication by the focal firm likely makes a minor difference. In nascent, under-explored areas, follow-on research is likely more valuable. Lastly, the availability of government funding to academics could

the distribution of resources in their favor.

also have an influence. An abundance of government funding increases the public good nature of science (Babina et al., 2023). Therefore, it likely reduces barriers for the originating firms in assimilating it into subsequent innovation.

3 Data

3.1 Sample Construction

I combine data from several sources: (i) corporate scientific publications, patents, and accounting information from the Duke Innovation & SCientific Enterprises Research Network (DISCERN, Arora et al. (2020)); (ii) scientific publications and citations data from Microsoft Academic Graph (Sinha et al., 2015); additional publications and disambiguated author data from Dimensions.ai; (iii) patent citations to scientific publications from the Reliance on Science in Patenting project (Marx & Fuegi, 2022); (iv) data on renowned scientists from the American Men and Women of Science (AMWS) directory; and (v) patent quality measures from Kelly et al. (2021) and Kogan et al. (2017).

The DISCERN dataset includes 582,107 journal articles and conference proceedings from Web of Science (WoS) associated with U.S.-based publicly-traded firms, published between 1980 and 2015. Based on these data, I create a crosswalk between WoS, Dimensions, and MAG that includes 471,153 records of firms' publications.[11] Second, I limit the sample to the years 1990-2012 due to truncation, missing values, and data irregularities in the earlier sample years. Third, the instrumental variable's logic requires using standard journal articles. Therefore, I remove conference proceedings and publications in special issues from the sample used in the main analysis.[12] Lastly, I remove observations with missing data (e.g., no journal volume or issue information) and singletons that result from the inclusion of fixed effects. My final sample at the publication level consists of 164,495 observations (156,475 unique publications) matched to 1,527 firms.

[11] 463,027 unique publications, since some publications are coauthored by researchers from multiple firms.

[12] See Section 5.1 for details. I present an analysis of conference proceedings in Appendix Section C.2.3.

3.2 Variables and Measures

See Appendix B for additional details regarding sources and data construction.

3.2.1 Explanatory Variable

I measure follow-on research by observing citations from outside the firm to the focal publication. I count up to three generations of citations (the first generation being direct citations). In cases where several routes connect the focal publication to the citing paper, I count the shortest route.[13] Most citations to firms' publications originate from academics employed by universities and other public research institutions. However, I also include citations from researchers at other firms.[14]

3.2.2 Dependent Variables

I am interested in identifying the effect of follow-on research on the originating firm's subsequent scientific and innovative performance. To explore the effects on scientific production, I test how follow-on research affects the scientific activities of the corporate authors of the focal paper. I provide three related measures. First, I count all corporate authors' scientific publications published after the focal paper. Next, I count the number of corporate authors' publications that involved collaborations with university academics (known as University-Industry Collaborations, UIC). Lastly, I count their conference proceedings as an indicator of academic conference participation.

Second, I estimate the effect of follow-on research on the subsequent hiring of scientists. I look for AMWS scientists hired by the firm *after* the focal publication year whose work is related to the focal publication. To determine relevance, I obtained a list of concepts extracted from the publication text for each focal publication.[15] Next, I look for publications by AMWS scientists and their associated concepts. If there is an overlap of concepts, I consider the scientist relevant. The benefit of this approach is that I obtain accurate employment years for these individuals. The challenge is that the data focus on a select set of scientists. Overall, I identified the employment

[13] See Appendix Figure B1 for an illustration.

[14] See Appendix Section C.2.5 for a split analysis.

[15] Concepts are extracted by Dimensions from titles and abstracts using a pointwise mutual information algorithm. See Appendix B for details.

of 20,552 individuals by DISCERN firms, of which I matched 6,673 to publications and for whom I could extract related concepts. Therefore, this measure is likely an underestimate of total hiring activities.

Lastly, I explore the effect follow-on research has on patenting activity using two approaches. First, I count the number of patents by the focal paper's authors assigned to the originating firm. I consider both a count of patents filed at least three years after the focal publication year and a count of patents that are filed at least ten years after. Second, I count the number of the firm's patents that cite the focal publication in a non-patent literature (NPL) citation. In addition, I count the firm's patents that cite either the focal publication or the follow-on research in an NPL citation.[16]

3.2.3 Control Variables

I view the number of citations a publication receives as a function of four variables. First, citations are affected by the scientific content of the publication. Second, the journal and publication year. Next, the authors' identity, since scientific prominence drives more attention to authors' works. Lastly, citations are affected by peer effects across authors whose publications appear in the same journal issue (see Section 5.1 for details).

I proxy for the prominence of authors by calculating their H-index at the time of publication. The H-index is a widely popular metric of the impact and productivity of academics.[17] I include the focal paper's author's H-index to control for their prominence. In cases where the focal publication has multiple authors, I use the highest H-index among the authors as the control variable. As I discuss in Section 5, I use the sum of the two highest H-indexes among all the other authors in the

[16]Patent citations to scientific articles typically reflect knowledge flows and the use of science in the invention process (Roach & Cohen, 2012). One complication, however, is that a count of patents that cite follow-on research is potentially mechanically related to the count of follow-on publications that are available to cite. In other words, it is possible that the firm would have had the same number of related patents regardless of the citation count to follow-on research. With the caveat that a positive relationship can be at least partially driven by increased "exposure" to follow-on publications, this measure provides further evidence for the relation between follow-on research and internal use of that research in subsequent inventive activity.

[17]To calculate this measure, I identify all papers published by the author prior to publication year t and count all citations to these publications received up to that year. Next, I sort the papers by descending order of citation counts. Then, the H-index is defined as $h = \max\{i \in \mathbb{N} : f(i) \geq i\}$, where $f(i)$ is the citation count for the publication in position i. For example, if an author has five publications with citation counts 33, 20, 8, 4, and 1, their H-index would be 4, as the publication at the 4th position has at least four citations.

same journal issue as the focal publication as an instrumental variable that is arguably exogenous to the scientific content of the focal publication, after controlling for differences across journals and years.

I control for differences between journals at different points in time by including a complete set of journal-year fixed effects. Lastly, I include a set of firm fixed effects to control for time-invariant differences across firms.

4 Descriptive Analysis

A narrow focus on within-firm direct patent citations provides a limited view that might suggest a disconnect between firms' scientific investments and patenting activities. Through tracking patent citations to external follow-on research, I reveal a contrasting and more comprehensive view that supports a strong relationship between science and invention within the firm. Table 1 classifies corporate publications into three groups.[18] The first group includes publications cited by a patent within the firm, suggesting direct use of the underlying science in a firm's invention. The second group contains publications that are not directly cited by an internal patent. Instead, for these publications, there is a patent by the firm that cites their external follow-on research. While not used directly, these publications are tied indirectly to the firm's inventions through follow-on science. The third group includes publications not used (directly or indirectly) by the firm's patents.

After accounting for truncation, I find that only about 7% of firms' publications are directly tied to firms' patents. However, once indirect citations are observed, I find that about 40% of the publications are tied to patents through direct or indirect citations (after accounting for truncation). This figure suggests a close connection between science and invention within the firm.[19]

Figure 1 explores the timing of direct and indirect patent citations. Direct citations mainly

[18] For this analysis, I use the complete sample of DISCERN publications with a MAG identifier.

[19] Appendix Section C.1.1 presents a similar result from the direction of patents. After accounting for truncation, I find that more than 40% of corporate science-based patents are directly or indirectly related to firms' contributions to public research.

Table 1: Firms' Scientific Publications Cited by Own Patents

	All Publications		Years 1980-2000	
	Count	Percent	Count	Percent
Firm's patent directly cites paper	$26,744$	5.68%	$15,718$	6.99%
Firm's patent cites follow-on research				
1st Generation	$24,120$	5.12%	$16,307$	7.25%
2nd Generation	$38,508$	8.17%	$27,758$	12.34%
3rd Generation	$42,733$	9.07%	$30,534$	13.58%
All	$105,361$	22.36%	$74,599$	33.17%
Cited by a firm's patent (directly or indirectly)	$132,105$	20.04%	$90,317$	40.16%
Not cited by a firm's patent	$339,048$	71.96%	$134,564$	59.84%
Total	$471,153$	100.00%	$224,881$	100.00%

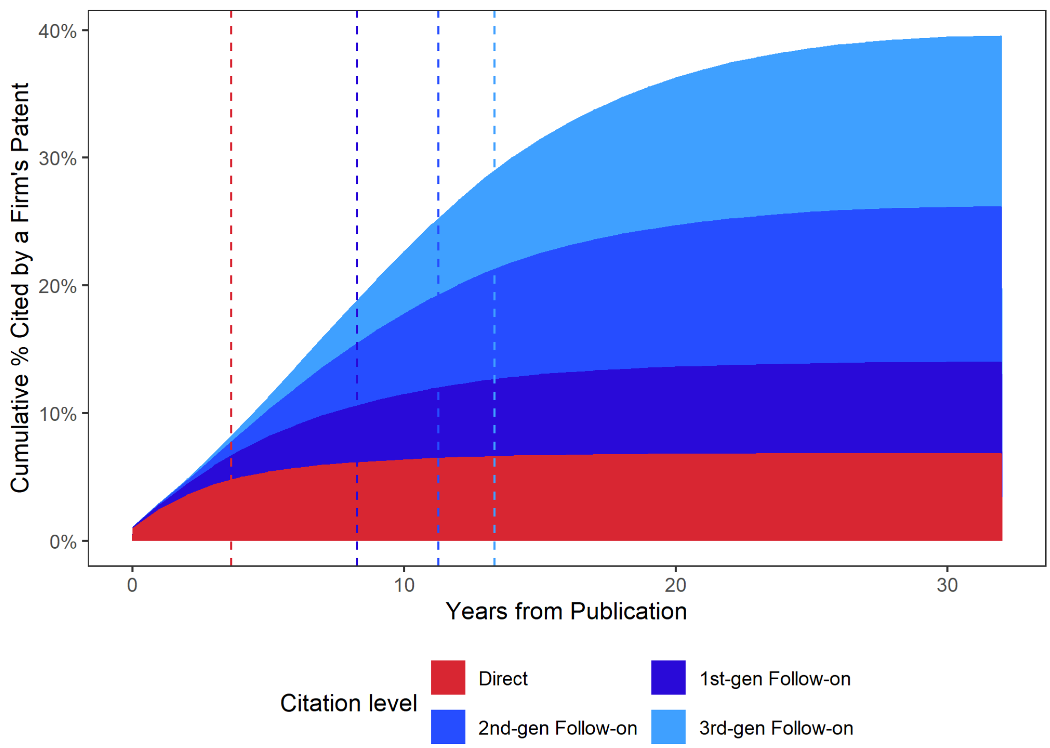
This table presents a summary of corporate scientific publications cited by the firms' own patents (NPL). The dataset includes 471,153 scientific publications. A publication is categorized by the shortest route to a citation by a firm's patent. 27,758 (6%) publications are directly cited by a patent of the same firm. An additional 105,361 (22%) publications are cited by external scientific publications that are subsequently cited by a firm's patent. Citations are followed up to the 3rd generation. 339,048 publications (72%) are not cited (directly or indirectly) by the firm's patents. To account for truncation, columns 3 and 4 present a subset of publications that are published until the year 2000. Among these publications, about 40% are eventually cited by the firms' patents.

occur within a few years of the focal paper publication year. The average direct patent citation occurs within 3.5 years of publication. Citations to follow-on research (indirect citations) are typically further out as follow-on research accumulates and becomes available. The average times for indirect citations are 8.2, 11.2, and 13.4 years for first-, second-, and third-generation follow-on, respectively.

5 The Effect of External Follow-On Research on Firms' Innovation

Identifying the relationship between external follow-on research and a focal firm's innovation outcomes is complicated by the potential endogeneity caused by the unobserved quality of the underlying science. For example, compared to less valuable research, scientifically important publications are more likely to receive citations from the scientific community and lead to subsequent internal investments by the firm. To address this issue, I propose an instrumental variable that can arguably control for this source of endogeneity. The instrumental variable uses a potential exogenous source of variation based on publishers' quasi-random grouping of accepted manuscripts into

Figure 1: Direct and Indirect Patent Citations to Firms' Publications



Note: This figure shows the cumulative percentage of corporate publications cited by the originating firms' patents. Direct citations are patents by the originating firm that directly cite the publication. Citations to follow-on research are patent citations to external follow-on research that cites the focal publication. A patent is counted once for each focal publication based on the shortest citation route. Citation timing is based on patent filing years. The dotted lines represent the average times to citations.

journal issues and the resulting shifts in scholars' attention to the focal publications.

5.1 Identification Strategy

Academic publishers typically follow a first-in-first-out principle to group accepted manuscripts into journal issues. At least until the early 2000s, when academic readership moved online, academic attention to a journal's issues varied with the prominence of authors included in that issue. Heightened attention to an issue could draw more attention to a focal publication, irrespective of its scientific content. This increased attention could have led to more follow-on research and related findings, manifesting as more citations to the focal publication. I rely on this quasi-random process and develop a measure of scholarly attention based on the H-index of the top two authors in the same journal issue as the focal publication. Below, I discuss the details of this process and the required assumptions for the instrument's validity.[20]

5.1.1 The Allocation of Accepted Manuscripts to Journal Issues

After a scientific manuscript is submitted to a journal, it undergoes peer review by highly trained academic professionals who assess its quality and relevance for publication. If the manuscript is accepted, it moves to the editorial staff, who prepare it for publication in the journal. The staff groups the accepted manuscripts into issues, which are typically published at a monthly, bi-monthly, or quarterly rate. It is worth noting that the editorial staff are not academic professionals, and the grouping of manuscripts into issues is not related to the content of the manuscripts.[21] Since the grouping of manuscripts into issues is primarily a result of the chronological order of acceptance, it can be considered a quasi-random process.[22] Therefore, in general, given a journal's acceptance of a manuscript, it is not possible to infer the quality of a publication through its allocation into adjacent journal issues.

5.1.2 Variation in Scholars' Attention to Journal Issues

Attention is a limited resource (Ocasio, 1997). Several recent works have highlighted how the competition for attention could affect the direction of scientific research and innovation (Bikard & Marx, 2020; Chai & Menon, 2019; Simcoe & Waguespack, 2011). Before the rise of the internet and the shift to online reading in the early 2000s, researchers primarily accessed scientific publications by physically going to their institution's library and borrowing journal issues. The scientific community's attention to issues of the same journal varied depending on their content. Publications by prominent authors drew more attention to specific journal issues than other issues of the same journal without such publications. As a result, adjacent publications in these journal issues received more exposure than those in other issues of the same journal. Therefore, the level of attention to journal publications could have varied regardless of their content or underlying quality.

[20] See Appendix Section D for a detailed discussion and validation tests.
[21] An exception is special issues and thematic journal issues, which are excluded from my main analysis.
[22] These details have been confirmed by executives at leading scientific publishing firms.

Increased exposure and scholarly attention to scientific publications can result in more follow-on research that relates to these publications. Exposure can generate new ideas and knowledge combinations and influence academics to work on problems they have not previously considered. Some prior evidence is available for this mechanism. Hudson (2007) studied citations to papers published in top economics journals. They found that the "traffic" an issue receives (total number of citations to publications in an issue excluding the focal publication) is positively correlated with citations to the focal publication. The inclusion of a "major paper" in an issue (a paper with citations over a given threshold) also predicts the number of citations to a focal publication. More recently, Lund and Maurya (2020) used data on publications in information science journals and found a positive relationship between citations to "highly-cited" papers and citations to other publications in the same journal issues.

5.1.3 Serendipity in Information Seeking

Nowadays, precise information retrieval is simple and rapid. However, before the rise of online academic readership and efficient search engines, academics relied on journal reading and browsing to keep up with relevant literature. Similarly to other aspects of scientific research, exposure to relevant information required some level of serendipity (Erdelez, 1999; Foster & Ford, 2003). Unanticipated encounters with information through various channels have often led to valuable outcomes (Makri & Blandford, 2012). The physical structure of journals exposed academics to information, sometimes without their previous intent. Journal browsing allowed scholars to identify new research ideas and direct their subsequent work.

Therefore, the key assumption for the instrument's relevance is that the prominence of other authors in a journal issue would drive more attention to the focal publication than publications in other issues of the same journal. This increased attention would translate into more follow-on research that would eventually be useful for the originating firm in subsequent innovation.

5.1.4 Instrument Specification

To proxy for increased attention, I calculate the H-index of all authors of publications in the same journal issue of each focal publication in the data. To ensure that further scientific developments do not affect the H-index, I calculate it based on the year before the focal publication year. Next, I identify the top two authors by H-index for each publication and use the sum as the instrument. Appendix Table D1 reports first-stage regressions (column 2) and several alternative specifications.

The central assumption for the instrument's validity is that the prominence of other authors does not confound with the quality of the focal publication's content and its authors' prominence. For this assumption to hold, the key is to restrict the comparison within journals and a reasonable time window. Therefore, I include a set of journal-year fixed effects. Conditional on these fixed effects, the instrument is plausibly unconfounded with respect to the outcome variables of interest. I provide several validity tests for these assumptions in Appendix D. First, I show that the instrument is uncorrelated with the H-index of the focal authors. Otherwise, it would have indicated that prominent authors tend to group into specific issues, even within a journal year. Next, I provide a placebo test that replaces the instrument with the H-index of the top authors from a random issue in that journal and year. The placebo instrument does not predict the count of follow-on research to the focal publication. Lastly, I show that the instrument had a stronger effect in the years before the early 2000s, when online readership replaced physical journal issues.

5.1.5 Comparing OLS and 2SLS Estimations

It is important to note that, on average, citations induced by the instrument could differ from those generated regardless of it. Possibly, "marginal" citations due to luck and serendipity might be less likely to represent scientific advancements that significantly build on the focal publication. Such citations could merely reflect awareness by others and the relevance of the focal publication to ongoing external scientific inquiry. These citations would less likely indicate novel follow-on findings that the firm can subsequently use as inputs. Instead, they will validate and increase the focal authors' scientific prominence. In this case, estimates from instrumented regressions could differ substantially from single-stage OLS regressions (see Appendix Section D.5 for details). Under these assumptions, OLS will recover correlations that average both types of citations, while 2SLS will put more weight on the role of follow-on research as validation.

A second potential source of difference between the OLS and 2SLS estimates is treatment heterogeneity. The effect of the instrument on follow-on research can vary based on the levels of prominence of the focal authors. Prominent focal authors will likely be less affected by neighboring authors in the same journal issue than less renowned authors. I find evidence for heterogeneity in treatment in Appendix Figure D5. Therefore, the 2SLS estimation could reflect the local effect on a subset of authors within the sample.[23]

5.2 Baseline Estimation Results

The econometric specification at the publication level (i) is as follows.

In the first stage, I estimate:

$$\ln(\text{Follow-On Research})_i = \alpha_1 \text{IV}_i + \alpha_2 \ln(\text{Focal H-index}_i) + \eta_f + \tau_t \times \gamma_j + \epsilon_i \tag{1}$$

The second stage is:

$$Y_i = \beta_1 \log(\widehat{\text{Follow-On Research}})_i + \beta_2 \ln(\text{Focal H-index}_i) + \eta_f + \tau_t \times \gamma_j + \epsilon_i \tag{2}$$

In the first stage, follow-on research is the count of three generations of external citations to the focal publication. The instrument is the sum of the two highest H-indexes of other authors in the same journal issue as the focal publication, a proxy for the academic attention to the journal issue that includes the focal publication. In the second stage, the dependent variable is the outcome of interest. All models include a control for the highest H-index among the authors of the focal publication, and a set of firm fixed-effects controls for time-invariant differences across firms. Journal-year fixed effects are required to establish the unconfoundedness of the instrument. Standard-errors are clustered by firm.

Econometric models with count-dependent variables are typically estimated using PPML or

[23]See Appendix Section D.4 for details.

OLS with log-linear specifications. For OLS models, a challenge arises when zero counts are present. A "popular fix" of adding a constant (usually 1) makes the log transformation feasible. This procedure is widely popular and, unfortunately, arbitrarily biased by the size of the chosen constant. This routine "produces estimates that lack meaningful interpretation and suffer from inherent biases that can cause them to have the wrong sign in expectation" (Cohn et al., 2022). A better solution is to use Poisson regressions. However, there is currently no accepted implementation of a two-staged Poisson regression with fixed effects. Therefore, in the main analysis, I report estimations of OLS and 2SLS regressions of linear probability models, where the dependent variable is equal to one if the count is greater than zero. I report corresponding single-staged Poisson estimates of count models in Appendix C.

Table 2 provides summary statistics for the publication sample.

Table 2: Descriptive Statistics for Publication Sample

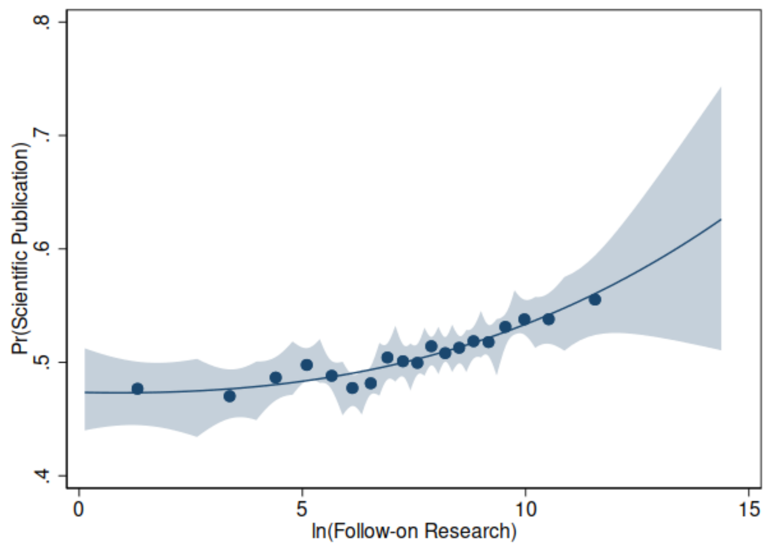
Variable	Mean	SD	Min	p25	p50	p75	Max
Publication Year	2,002.0	6.4	1990	1997	2002	2008	2012
Follow-On Research	12,932.3	39,453.0	1	355	2083	9280	2102260
Focal H-Index	19.3	17.8	0	7	15	27	173
Top Two Researchers H-Index (IV)	86.3	48.2	3	50	76	114	420
Future Pubs	18.2	41.2	0	0	1	17	933
Future UIC Pubs	7.3	21.0	0	0	0	5	858
Future Conf. Proceedings	1.0	6.6	0	0	0	0	332
AMWS Hires	0.3	1.8	0	0	0	0	58
AMWS Hires (Award-Winning)	0.0	0.2	0	0	0	0	6
Future Patents by Authors	10.0	30.5	0	0	1	7	1388
Future Patents by Authors (gap $\geq$ 10y)	1.8	11.1	0	0	0	0	830
Future Patents by Authors (UIC)	0.2	2.0	0	0	0	0	151
Int. Patents Citing Focal	0.3	7.1	0	0	0	0	1888
Int. Patents Citing Focal or FO	12.7	90.0	0	0	0	1	3819
Ext. Patents Citing Focal	2.3	18.5	0	0	0	1	3549
Patent-Paper Pair	0.1	0.3	0	0	0	0	1
University-Industry Collaboration	0.6	0.5	0	0	1	1	1
Low Concept Prevalence	0.6	0.5	0	0	1	1	1
High Government Funding	0.4	0.5	0	0	0	1	1
Future Patents by Authors (Citing FO)	0.3	4.4	0	0	0	0	467
Future Patents by Authors (Not Citing FO)	9.7	29.3	0	0	1	7	1388

Notes: This table provides summary statistics for the variables used in the econometric analysis at the publication level. The data is based on the DISCERN database of publications by U.S.-based publicly-owned firms between 1990 and 2012. The sample includes 164,495 publications originating from 1,527 firms.

5.2.1 The Effect of Follow-On Research on Firms' Investments in Science

Table 3 presents an estimation of the effects of follow-on research on firms' subsequent investments in science. I consider outcomes relating to subsequent scientific publications by the focal publications' corporate authors and the hiring of scientists whose work is closely related to the

Figure 2: Follow-On Research and Subsequent Scientific Publication



Note: This figure presents a binned scatterplot of the relation between logged follow-on citations and the probability of subsequent scientific publishing by the focal authors. The values are fitted by controlling for the logged H-index of the focal author, firm fixed effects, and journal-year fixed effects.

focal publication. First, I find a positive effect of follow-on research on the probability of subsequent publications by the focal authors (columns 1-2).[24] Figure 2 presents a binned scatterplot corresponding to column 1. Second, I find a positive effect of follow-on research on subsequent research collaborations with academics (UIC) and on academic conference participation by the focal authors (columns 3-6).[25] Table C2 column 3 provides evidence that the positive association remains even after controlling for the total number of future publications, suggesting that follow-on research is associated with a larger share of UIC among the authors' future research. In Table 3 column 5, I find a positive association between follow-on research and the probability of a

[24]According to the OLS estimate, for the average publication, a 1% increase in follow-on research is associated with $\frac{0.008/100}{0.507} = 0.016\%$ increase in the probability of at least one additional publication (column 1). The estimate from the corresponding 2SLS model is higher, suggesting a 1% increase in follow-on accounts for $\frac{0.121/100}{0.507} = 0.24\%$ increase in the probability of a subsequent publication (column 2). According to a corresponding count model estimated using Poisson regression, for the average observation, an additional publication by the focal authors requires a 90% increase ($(20.94 \times 0.053)^{-1}$) in the amount of follow-on research (Appendix Table C2).

[25]According to column 3, a 1% increase in follow-on research is associated with a 0.025% increase in the probability of a subsequent UIC publication. The 2SLS estimate in column 4 suggests a 0.28% local effect. Based on a Poisson regression, for the average observation, an additional UIC publication by the focal authors is associated with a 170% increase in the amount of follow-on research (Appendix Table C2, column 2).

subsequent conference proceeding by the focal authors.[26] Third, I explore the effect of follow-on research on the focal firm's subsequent hiring of renowned scientists. In Table 3 columns 7-8, I find evidence that follow-on research increases the probability that the firm will hire a scientist whose research is highly-related to the content of the focal publication. In columns 9-10, I find that these results hold even when focusing on award-winning scientists.[27] The estimated effects, however, are relatively small, possibly due to the partial availability of data regarding the hiring of scientists by the firms in the sample.

5.2.2 The Effect of Follow-On Research on Firms' Patenting Outcomes

Table 4 presents an estimation of the effects of follow-on research on firms' patenting outcomes. I consider two sets of outcomes. The first outcome of interest is subsequent patents that list the corporate authors of the focal paper as inventors and are assigned to the focal firm (columns 1-6). When considering future patents filed at least three years after the focal publication year, the OLS estimate is positive and statistically significant. In the instrumented model (column 2), this relation is positive but not statistically significant ($p = 0.11$). Figure 3 presents a corresponding binned scatterplot to column 1. However, when considering patents filed after a longer time gap after the focal publication year (ten years or more), there is a positive and statistically significant coefficient estimate both for the OLS and 2SLS models (columns 3-4).[28] This result makes sense, as follow-on research requires a long time to materialize and become available for the firm. A small subset of subsequent patents are assigned to both the focal firm and a public research institution. In column 5, I find a positive relation between follow-on research and a subsequent collaboration in patenting.[29]

The second set of outcomes in Table 4 includes patents by the firm that cite the focal publication and patents that cite either the focal publication or the follow-on research. I find evidence

[26]The 2SLS estimate in column 6 is positive but is not statistically significant ($p = 0.11$). I also find a corresponding positive relationship in the Poisson estimate in column 4 of Appendix Table C2.

[27]Corresponding estimates from count models (Appendix Table C2 columns 5-6) are also positive.

[28]Column 4 suggests that a 1% increase in follow-on research is associated with a 0.6% increase in the probability of at least one patent. Similar results are available for five- and seven-year minimum gaps.

[29]This result persists in a Poisson estimation of a count model (Appendix Table C3 column 3) but is too weak for the 2SLS estimation (column 6).

Table 3: The Effect of Follow-On Research on Firms' Investments in Science

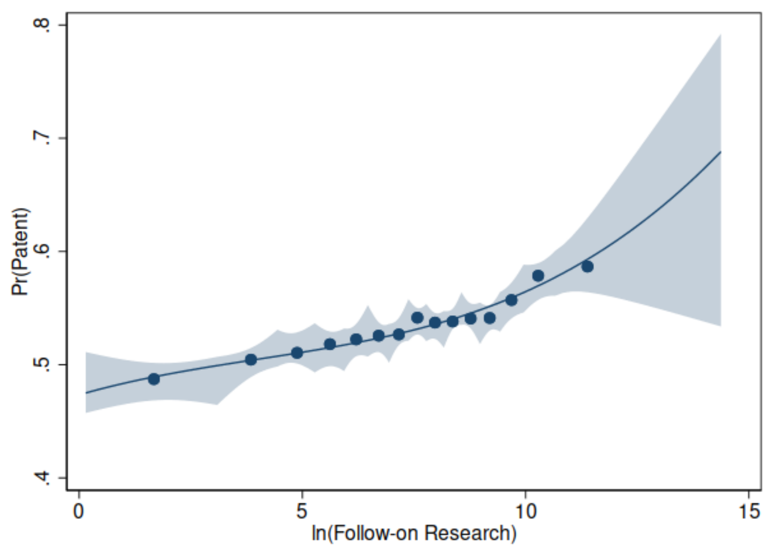
	Subsequent Scientific Publications by Focal Authors						Hiring of Renowned Scientists (AMWS)			
	Pr(Publication)		Pr(Univ. Collab.)		Pr(Conference Proc.)		Pr(Hire)		Pr(Award-Winning Hire)	
	OLS	2SLS	OLS	2SLS	OLS	2SLS	OLS	2SLS	OLS	2SLS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
ln(Follow-On)	0.008***	0.121**	0.011***	0.132***	0.002***	0.047	0.005***	0.053**	0.002*	0.035**
	(0.001)	(0.047)	(0.002)	(0.048)	(0.001)	(0.030)	(0.001)	(0.025)	(0.001)	(0.015)
ln(Focal H-Index)	0.002	-0.030**	0.035***	0.001	0.005**	-0.008	0.003**	-0.011	0.001	-0.009**
	(0.003)	(0.014)	(0.003)	(0.014)	(0.002)	(0.009)	(0.001)	(0.007)	(0.001)	(0.004)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Journal-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	164,495	164,495	164,495	164,495	164,495	164,495	164,495	164,495	164,495	164,495
Avg. DV	0.507	0.507	0.435	0.435	0.094	0.094	0.080	0.080	0.018	0.018
First Stage F-stat		29.528		29.528		29.528		29.528		29.528
Adjusted R^2	0.344	-0.377	0.320	-0.398	0.257	-0.259	0.223	-0.289	0.210	-0.409

Notes: This table presents estimation results for the relationship between external follow-on research and firms' subsequent investments in science. The data consists of a pooled cross section of publications by U.S.-based publicly-owned firms, published between 1990 and 2012 (Arora et al., 2020). Follow-on research is the total count of three generations of citations to the focal publication from outside the firm. The dependent variable is an indicator for future scientific publications (columns 1-2), future publications with academics (columns 3-4), and future conference proceedings (columns 5-6) written by the corporate authors of the focal publication. In columns 7-8, the dependent variable is an indicator for future employment of a renowned scientist (columns 9-10, award-winning scientist) whose work is highly related to the focal publication. In 2SLS regressions, the instrumental variable is the sum of H-indexes of the top two authors in the same journal issue as the focal publication. All regressions include a control for the highest H-index among the authors of the focal publication, as well as firm and journal-year fixed effects.
Clustered (Firm) standard-errors in parentheses. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

that follow-on research increases the probability of a patent by the firm citing the focal publication (columns 7-8).[30] In columns 9-10, I consider patent citations to both the focal and the follow-on research. I find additional positive effects, with the caveat that this relation is partly mechanical.[31] These results indicate a positive effect of follow-on research on subsequent patenting outcomes by the originating firms.

Taking together the results in Tables 3 and 4, I find evidence that follow-on research increases firms' subsequent investments in science, improves their patenting outcomes, and drives direct ties with academics.

Figure 3: Follow-On Research and Subsequent Patenting



Note: This figure presents a binned scatterplot of the relation between logged follow-on citations and the probability of subsequent patenting by the focal authors. The values are fitted by controlling for the logged H-index of the focal author, firm fixed effects, and journal-year fixed effects.

[30] According to column 7, for the average publication, a 1% increase in follow-on research is associated with a 0.2% increase in the probability of a citing patent by the firm. The instrumented model in column 8 suggests a positive effect of 0.75% on the probability. A corresponding estimate from a count model (Appendix Table C3 column 4) suggests that for the average observation, an additional citing patent within the firm is associated with 3.75 times the amount of follow-on research.

[31] More follow-on research provides more opportunities for citations. For the average observation, a 1% increase in follow-on research is associated with a 0.4% increase in the probability of at least one citation (column 10). According to a corresponding count model, an 8% increase in follow-on is associated with an additional patent citation by the firm (Appendix Table C3 column 5).

Table 4: The Effect of Follow-On Research on Firms' Patenting Outcomes

	Subsequent Patents by Focal Authors						Subsequent Firms' Patents			
	Pr(Patent, ≥3y gap)		Pr(Patent, ≥10y gap)		Pr(Patent, Univ. Collab.)		Pr(Citation to Focal)		Pr(Citation to Focal or FO)	
	OLS (1)	2SLS (2)	OLS (3)	2SLS (4)	OLS (5)	2SLS (6)	OLS (7)	2SLS (8)	OLS (9)	2SLS (10)
ln(Follow-On)	0.007***	0.072	0.003***	0.095**	0.002***	0.006	0.011***	0.043*	0.074***	0.128***
	(0.001)	(0.046)	(0.001)	(0.040)	(0.000)	(0.018)	(0.001)	(0.026)	(0.004)	(0.039)
ln(Focal H-Index)	0.012***	-0.007	0.002	-0.024**	0.009***	0.008	-0.008***	-0.017**	0.002	-0.013
	(0.002)	(0.013)	(0.002)	(0.011)	(0.002)	(0.005)	(0.001)	(0.008)	(0.002)	(0.012)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Journal-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	164,495	164,495	164,495	164,495	164,495	164,495	164,495	164,495	164,495	164,495
Avg. DV	0.397	0.397	0.143	0.143	0.033	0.033	0.056	0.056	0.303	0.303
First Stage F-stat		29.528		29.528		29.528		29.528		29.528
Adjusted R^2	0.348	-0.248	0.351	-0.486	0.128	-0.170	0.071	-0.224	0.457	-0.109

Notes: This table presents estimation results for the relationship between external follow-on research and firms' subsequent patenting outcomes. The data consists of a pooled cross section of publications by U.S.-based publicly-owned firms, published between 1990 and 2012 (Arora et al., 2020). Follow-on research is the total count of three generations of citations to the focal publication from outside the firm. The dependent variables are indicators for a future patent filed at least three years after the publication of the focal paper (columns 1-2), at least ten years after the publication of the focal paper (columns 3-4), and a future patent that is originally assigned to both the firm and a public research institution (columns 5-6). In columns 7-8, the dependent variable is an indicator for a future patent by the firm that cites the focal publication (columns 9-10, the focal publication or the follow-on research). In 2SLS regressions, the instrumental variable is the sum of H-indexes of the top two authors in the same journal issue as the focal publication. All regressions include a control for the highest H-index among the authors of the focal publication, as well as firm and journal-year fixed effects.
Clustered (Firm) standard-errors in parentheses. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

5.2.3 Matched Controls

One possible concern is that the firm's involvement in scientific research is inconsequential. That is, possibly the effect of follow-on research would be similar to completely external advancements in a given scientific research topic. While I cannot observe the counterfactual for firms' investments in research, Table 5 provides a step in that direction. For each focal ("internal") publication in my sample, I match a random publication that is not associated with the focal firm and published in the same journal and year. Next, I observe the focal firm's patent citations to both the original sample and the sample of matched publications. The dependent variable is an indicator equal to one if at least one patent by the focal firm cites the internal (or matched) publication. As expected, I find that the estimated coefficient for internal publications is larger than the estimated coefficient for the matched publication sample. Columns 4-5 estimate an interaction term of log(follow-on) with an indicator for an internal publication. I find a positive and significant interaction term using OLS (column 4) and a weakly significant estimate using 2SLS (column 5, p=0.09).[32] These results suggest that follow-on research is more strongly associated with a subsequent patent when firms rely on their own science. Of course, these results should be interpreted with caution, as firms choose the topics of their research.

5.3 Mechanisms Driving the Value of Follow-On Research

Follow-on research can be valuable for firms through multiple mechanisms. First, it can provide valuable inputs that the originating firms can incorporate into subsequent innovation. Firms can incorporate such inputs either by leaning on internal scientific capabilities and absorbing external knowledge or by setting up direct ties with external researchers. Second, follow-on citations that do not represent useful inputs can benefit the originating firms as quality validation for the internally-produced scientific findings. Such validation could influence internal resource allocation and impact the direction of innovation within the firm.

Studying the coefficient estimates on the effects of follow-on research on subsequent scientific

[32]Estimated for the average publication, the probability of a patent citation to the internal publication following an increase of 1% in follow-on research is 2.5 times ($\frac{0.008+0.045+0.02}{0.029}$) the probability of a patent citation to the matched publication (following a similar increase).

Table 5: Comparison of Internal Publications and Matched Controls

	Pr(Subsequent Firms' Patents Citing Focal Publication)				
	OLS Internal Pubs	OLS Matched Pubs	OLS	OLS	2SLS
	(1)	(2)	(3)	(4)	(5)
ln(Follow-On) × Internal				0.011***	0.008*
				(0.001)	(0.005)
Internal			0.051***	0.049***	0.045***
			(0.004)	(0.004)	(0.005)
ln(Follow-On)	0.011***	0.001***		0.000	0.020
	(0.001)	(0.000)		(0.001)	(0.015)
ln(Focal H-Index)	-0.008***	-0.000	-0.003***	-0.004***	-0.010**
	(0.001)	(0.000)	(0.001)	(0.001)	(0.005)
Firm FE	Yes	Yes	Yes	Yes	Yes
Journal-Year FE	Yes	Yes	Yes	Yes	Yes
Observations	164,494	158,776	347,088	347,088	347,088
Avg. DV	0.056	0.004	0.029	0.029	0.029
First Stage F-stat					29.149
Adjusted R^2	-0.163	-0.171	-0.087	-0.074	-0.112

Notes: This table compares the sample of internal publications and a sample of matched publications originating outside the firm. For each internal publication, a publication from the same journal year is randomly assigned. The dependent variable is an indicator variable that is equal to one if at least one patent within the firm is citing the publication. Follow-on research is the total count of three generations of citations to the focal publication from outside the firm. In interaction models (columns 4-6), ln(follow-on) is recentered around the sample mean. Internal is an indicator variable that is equal to one if the publication is published by the originating firm and equal to zero if it is the matched external publication. In 2SLS regressions, the instrumental variable is the sum of H-indexes of the top two authors in the same journal issue as the focal paper. Focal H-index is the highest H-index among the authors of the focal paper at the year of publication. In column 5, ln(follow-on) and the interaction term are instrumented by the the IV and the interaction of the IV with the indicator for internal publication. Observations are automatically dropped from the full sample due to separation and singletons.

Clustered (Firm) standard-errors in parentheses. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

investments (Table 3) and patenting outcomes (Table 4) reveals that the OLS estimates tend to be smaller than their 2SLS counterparts. These relations might be counterintuitive, given the expected positive bias of the OLS due to the unobserved scientific quality of the focal publications. However, the larger 2SLS estimates can potentially indicate the mechanisms at play. Citations that are predicted by the instrument are more likely to provide validation rather than meaningful inputs. A large effect due to these citations could suggest that follow-on research is beneficial through the abovementioned roles.

To further explore these mechanisms, I develop measures for three ways firms can use follow-on research. First, I proxy for the use of follow-on research as an input into invention by observing patents by the focal authors that cite the follow-on research. Second, I proxy for university-industry collaborations in invention by observing patents that are co-assigned to the firm and a university. Third, I proxy for follow-on research that provides external validation by observing patents that do not cite the follow-on research. I view the lack of citation as an indicator that the research was not an input into invention. Next, I argue that when the authors of the focal publication are less prominent, firms face more uncertainty regarding their quality. In these cases, follow-on research is more likely to be useful for quality validation than when there is less uncertainty regarding the focal authors' quality of work. Moreover, for prominent focal authors, follow-on research is more likely to be beneficial as an input, either disembodied (as a citation) or embodied (through a collaboration).

Table 6 reports the estimation results. In general, while the estimates are noisy, they provide some evidence in support of the mechanisms discussed above. I interact the effect of follow-on research with an indicator for above-median focal author prominence (measured as the top H-index among the focal authors). The dependent variables are indicators for a subsequent patent by the focal authors, where patents are classified based on their citations to follow-on research and co-assignment to public research institutions. In columns 1-2, the dependent variable is the probability of any subsequent patent by the focal authors. When considering all patents, there is little evidence of a difference between low- and high-prominence focal authors. However, the differences are more pronounced when considering the different patent classifications. Follow-on research is more useful as an input when the focal authors have an above-median H-index (columns 3-6). In contrast, follow-on research is more useful as validation when the focal authors have a below-median H-index (columns 7-8). These results suggest that follow-on research can be valuable in multiple ways. The mechanisms that drive the private value from follow-on research depend on the firm's and authors' attributes.

Table 6: Ways in Which Follow-On Research Is Privately Useful

	Subsequent Patents by Focal Authors ($\geq$3 year gap)							
	Pr(Patent)		Pr(Patent, Citing FO)		Pr(Patent, Univ. Collab)		Pr(Patent, Not Citing FO)	
	OLS (1)	2SLS (2)	OLS (3)	2SLS (4)	OLS (5)	2SLS (6)	OLS (7)	2SLS (8)
ln(Follow-On) × High Focal H-Index	-0.000	-0.014**	0.003***	-0.002	0.002**	0.006**	-0.000	-0.016***
	(0.001)	(0.006)	(0.001)	(0.003)	(0.001)	(0.003)	(0.001)	(0.006)
High Focal H-Index	-0.002	-0.002	-0.002	-0.001	-0.001	-0.002	-0.003	-0.003
	(0.004)	(0.005)	(0.002)	(0.003)	(0.002)	(0.003)	(0.004)	(0.005)
ln(Follow-On)	0.007***	0.076*	0.013***	0.030	0.001***	0.005	0.007***	0.078*
	(0.001)	(0.046)	(0.002)	(0.022)	(0.000)	(0.018)	(0.001)	(0.046)
ln(Focal H-Index)	0.013***	-0.004	0.004***	-0.000	0.009***	0.008*	0.012***	-0.005
	(0.003)	(0.012)	(0.001)	(0.005)	(0.002)	(0.005)	(0.003)	(0.012)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Journal-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	164,495	164,495	164,495	164,495	164,495	164,495	164,495	164,495
Avg. DV	0.397	0.397	0.046	0.046	0.033	0.033	0.394	0.394
First Stage F-stat		14.416		14.416		14.416		14.416
Adjusted R^2	0.348	-0.246	0.142	-0.170	0.128	-0.172	0.345	-0.253

Notes: This table explores how the effect of external follow-on research and firms' subsequent patenting outcomes vary by the prominence of the focal authors (see Table 4 for baseline results). In all models, follow-on research is interacted with an indicator for above-average prominence of the top focal author (within each journal). Prominence is measured using H-index values at the time of publication. The dependent variables are indicators for subsequent patents by the focal authors, filed at least three years after the focal publication. In columns 1-2, all subsequent patents are considered. Next, the table classifies patents into three categories. In columns 3-4, the dependent variable is an indicator for a subsequent patent that cites the follow-on research. In columns 5-6, the dependent variable is an indicator for a subsequent patent that is originally assigned to both the focal firm and a public research institution. In columns 7-8, the dependent variable is an indicator for a subsequent patent that does not cite the follow-on research. In all models, log(follow-on) is recentered around the sample mean. In 2SLS models, ln(follow-on) and the interaction term are instrumented by the the IV and the interaction of the IV with the indicator variable.

Clustered (Firm) standard-errors in parentheses. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

5.4 Conditions That Drive the Private Value of Follow-On Research

5.4.1 Competition and Technological Leadership

Corporate publications create opportunities for the emergence of privately valuable follow-on research. However, they also increase the risk of rivals benefiting from the underlying scientific findings. As more external research follows a given focal publication, others will likely be more inclined to use it in their inventions, potentially to the detriment of the originating firm. Table 7 column 1 provides evidence that follow-on research positively correlates with external patent citations to the focal publication.[33] These findings suggest that external follow-on research can also increase the risk of negative knowledge spillovers from firms' publications.

The risk of negative knowledge spillovers can vary by the technological competition the focal firm faces. Technological leaders are likely to benefit more from follow-on research than firms facing fierce competition. When others' patents build on follow-on research, the focal firm will likely have less ability to appropriate the knowledge for its internal use. To study this relation, I develop a measure of firms' technological leadership at the publication level. Given the research fields of the publications in the sample, I identify a set of patent categories (CPC codes) most related to each publication.[34] I then identify related patents filed in the given year by the focal firm and patents by all firms in the sample. I use the percentage of patents by the focal firm as a measure of technological leadership. The larger this measure, the stronger the focal firm is in patenting related technologies compared to other firms. I use a cutoff at the sample median as an indicator for technological leadership at the focal publication level.

Table 7 columns 2-5 explore how the benefits of follow-on research vary with a granular measure of technological leadership. Note that the interaction term does not have a causal interpretation. However, results that are in line with the theory are a step toward explaining under which circumstances the effects may be stronger. In columns 2-3, the dependent variable is an indicator for a subsequent patent by the corporate authors of the focal paper. In an OLS esti-

[33]Note that the instrumental variable cannot be applied here, as external patent citations can themselves be a direct outcome of increased external attention.

[34]I use the complete NPL data from Marx and Fuegi (2022) to identify the top CPCs for each scientific research field. See Appendix B for details.

mation, I do not find evidence of a difference between the two subsamples. However, there is a positive and statistically-significant difference in the 2SLS estimation. In columns 4-5, the dependent variable indicates a patent by the originating firm that cites the focal publication. Here, the OLS estimation is positive and statistically significant. The 2SLS estimate, on the other hand, is positive but insignificant, possibly due to lack of power. These results provide some evidence that follow-on research is more strongly related to subsequent patenting in areas where firms are technological leaders. Overall, the results suggest that in areas where the focal firm has a leading patenting presence with respect to other firms, the benefits from follow-on research seem more significant.

Table 7: External Patent Citations and Technological Leadership

	Competition	Focal Firm			
	Pr(External Patents Citing Focal)	Pr(Subsequent Patent By Focal Authors)		Pr(Firms' Patent Citing Focal)	
	OLS (1)	OLS (2)	2SLS (3)	OLS (4)	2SLS (5)
ln(Follow-On) × Leader		0.002	0.049***	0.004***	0.001
		(0.002)	(0.014)	(0.001)	(0.005)
Leader		0.090***	0.083***	0.017***	0.018***
		(0.012)	(0.011)	(0.004)	(0.004)
ln(Follow-On)	0.050***	0.008***	-0.038	0.009***	0.042
	(0.002)	(0.001)	(0.049)	(0.001)	(0.026)
ln(Focal H-Index)	-0.007***	0.013***	0.019	-0.008***	-0.017**
	(0.002)	(0.002)	(0.014)	(0.001)	(0.008)
Firm FE	Yes	Yes	Yes	Yes	Yes
Journal-Year FE	Yes	Yes	Yes	Yes	Yes
Observations	164,495	164,495	164,495	164,495	164,495
Avg. DV	0.273	0.535	0.535	0.056	0.056
First Stage F-stat			14.594		14.594
Adjusted R^2	0.263	0.317	-0.186	0.072	-0.221

Notes: Column 1 presents estimation results for the relation between external follow-on research and external patent citations to the focal publication. Columns 2-5 explore how the baseline patenting results (presented in Table 4) vary with technological leadership. Leadership is defined as above-median patenting in related CPC categories in the year of publication across all firms in the sample. In columns 2-3, the dependent variable is an indicator variable for at least one future patent by the corporate authors of the focal paper. In columns 4-5, the dependent variable is an indicator variable for at least one patent by the firm that cites the focal paper. In 2SLS regressions, the instrumental variable is the sum of H-indexes of the top two authors in the same journal issue as the focal publication. Focal H-index is the highest H-index among the authors of the focal publication. Observations are automatically dropped from the full sample due to separation and singletons.
Clustered (Firm) standard-errors in parentheses. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

5.4.2 Firm Capabilities and Characteristics of Related Science

In addition to the moderation relations due to competition, the private value of follow-on research is also likely to be moderated by firms' scientific and inventive capabilities, and by characteristics of the external scientific community that are related to the focal publication. I explore such relations in Table 8. In columns 1-2, I explore the role of firms' possession of complementary intellectual property rights (IPR). To proxy for IPR, I search for patent-paper pairs, defined as patents by the focal authors filed at most two years after the focal publication that share textual concepts with the focal publication. To ensure that I do not include these patents in the construction of the dependent variable, I limit the outcomes to patents filed three years or more after the focal publication. I find a positive interaction term in both the OLS and 2SLS estimation, suggesting that complementary assets assist firms in subsequently benefiting from follow-on science.

Second, I explore the role of firms' decision to collaborate on the focal publication with external academics (UIC). On the one hand, external researchers can extend the firms' absorptive capacity and increase the use of follow-on research. On the other hand, outsourcing research through temporary collaborations could substitute for firms' internal scientific capabilities and reduce their ability to benefit from follow-on research in later years. In Table 8 columns 3 and 4, I find results that are in line with the latter. Compared to strictly in-house research, follow-on related to collaborated focal publications is less likely to lead to a subsequent patent by the focal authors.

Next, I contrast mature scientific areas with abundant related research to nascent areas where external exploration is scarce and limited. In areas where many academics are operating, the firm is potentially facing less need (or ability) to induce additional external inquiry. I construct a measure of scientific prevalence by counting the appearance of textual concepts related to the focal publications in previous publications outside the firm (up to three years apart). Columns 5 and 6 report the estimation results. Follow-on research is more valuable when it originates in areas less previously explored than in mature and well-developed areas. These results suggest that follow-on research is more likely to benefit the firm when it redirects academics to work on new and less-explored research areas.

Lastly, I compare scientific areas where government funding is relatively abundant to areas with less support from the government. According to Babina et al. (2023), government funding increases the public-good nature of science and makes it more available for appropriation (in this case, by the focal firm). Therefore, the availability of government funding will make follow-on research more valuable for the focal firm. I construct a measure of government funding availability by observing funding acknowledgments of prior research related to the focal publication (as identified in the previous results above). For each focal publication, I calculate the percentage of prior related works that acknowledge government funding and split the sample based on the median. Columns 7-8 report the results. I find evidence for a positive moderation of government funding availability.[35]

5.5 Additional Analyses

Appendix Section C.2 provides additional results. Namely, I explore heterogeneity analyses of the effect of follow-on research on subsequent patents by the firm citing the focal publication, correlations on a sample of firms' conference proceedings, variation of the baseline results by scientific fields and main industries, and a comparison of the effects of corporate versus academic follow-on research.

6 Follow-On Research and Firms' Innovative Performance

6.1 Follow-On Research and Patent Quality Measures

I study a pooled cross section of corporate patents and their citations to scientific publications (non-patent literature, NPL) and find that, on average, corporate patents that cite follow-on research tend to be of better quality and value than other science-based patents by the same firm. I document a positive relation between NPL citations to follow-on research and measures of patent value, legal scope, and textual novelty. These relations are comparable to those associated with

[35] Appendix Table C6 presents corresponding estimations for the effect on subsequent patents by the firm that cite the focal publication. In general, the results are qualitatively similar but with weaker statistical significance.

Table 8: Heterogeneity in Subsequent Patenting

	Firm Capabilities				Scientific Community			
	Pr(Subsequent Patent by Focal Authors, ≥3y gap)							
	OLS	2SLS	OLS	2SLS	OLS	2SLS	OLS	2SLS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Complementary IP Rights								
ln(Follow-On) × PPP	0.006***	0.024**						
	(0.002)	(0.011)						
PPP	0.304***	0.269***						
	(0.016)	(0.026)						
Knowledge Outsourcing								
ln(Follow-On) × UIC			-0.005***	-0.020***				
			(0.001)	(0.008)				
UIC			-0.066***	-0.064***				
			(0.006)	(0.005)				
Scientific Concept Prevalence								
ln(Follow-On) × Low Prev.					0.001	0.024***		
					(0.001)	(0.009)		
Low Prevalence					0.015***	0.008*		
					(0.003)	(0.005)		
Government Funding Availability								
ln(Follow-On) × Govt. Funding							0.003**	0.028*
							(0.001)	(0.016)
Govt. Funding							0.022***	0.005
							(0.004)	(0.009)
ln(Follow-On)	0.004***	0.081*	0.010***	0.077*	0.007***	0.060	0.006***	0.064
	(0.001)	(0.045)	(0.001)	(0.045)	(0.001)	(0.043)	(0.001)	(0.044)
ln(Focal H-Index)	0.010***	-0.012	0.023***	0.006	0.012***	-0.007	0.011***	-0.007
	(0.002)	(0.013)	(0.002)	(0.013)	(0.002)	(0.013)	(0.002)	(0.013)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Journal-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	164,495	164,495	164,495	164,495	164,495	164,495	164,495	164,495
Avg. DV	0.397	0.397	0.397	0.397	0.397	0.397	0.397	0.397
First Stage F-stat		15.215		14.452		14.423		14.904
Adjusted R^2	0.380	-0.233	0.352	-0.233	0.348	-0.257	0.349	-0.252

Notes: This table explores heterogeneity in the relationship between external follow-on research and firms' subsequent patenting outcomes (see Table 4 for baseline results). In columns 1-2, follow-on research is interacted with an indicator that the focal publication is a part of a patent-paper pair (PPP). In columns 3-4, follow-on research is interacted with an indicator that the focal publication is a university-industry collaboration (UIC). In columns 5-6, follow-on research is interacted with an indicator for below-average prevalence of prior works sharing the same textual concepts. In columns 7-8, follow-on research is interacted with an indicator for above-average government funding acknowledgments in related publications. In all models, log(follow-on) is recentered around the sample mean. In 2SLS models, ln(follow-on) and the interaction term are instrumented by the the IV and the interaction of the IV with the indicator variable.

Clustered (Firm) standard-errors in parentheses. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

NPL citations to internal research.

The econometric specification at the patent level (p) is as follows:

$$Y_p = \beta_1 \log(\text{NPL to Int})_p + \beta_2 \log(\text{NPL to FO})_p + \eta_f + \tau_t + \epsilon_p \tag{3}$$

The sample includes corporate patents from the DISCERN database with at least one NPL citation.[36] I distinguish between two types of NPL citations. NPL citations to internal publications (β_1) are direct citations to publications by the focal firm. NPL citations to follow-on research (β_2) are citations to external research with an upstream reference to a publication by the same firm (up to three generations away). I explore whether counts of these citations positively correlate with private value, legal scope, and textual novelty measures (Y_p). In all specifications, I include firm- and patent-grant-year fixed effects. Standard-errors are clustered by firm.

Table 9 presents the OLS estimation results. The estimated coefficients establish that patents' citations to follow-on research are positively related to higher private value,[37] broader legal scope,[38] and textual novelty.[39] In addition, the relation is comparable in magnitude to the relation of citations to internal publications. Lastly, I compare the coefficient estimates of citations to follow-on research and citations to internal publications. I find that NPL citations to internal publications have a stronger, but comparable, correlation with private value and scope. Taken together, these results provide correlational support at the patent level for the positive effects of follow-on research on firms' innovation outcomes.

[36] Descriptive statistics are provided in Table B2.

[37] According to column 1, a 1% increase in the citation count to follow-on research is associated with a 0.05% increase in value.

[38] According to column 3, a 1% increase in the citation count to follow-on research is associated with a reduction of 0.1% in the number of words. The fewer words in a claim description, the broader is the legal scope of the patent (Kuhn & Thompson, 2019).

[39] About 19% of the patents in the sample are considered breakthrough patents in terms of textual novelty (Kelly et al., 2021). According to column 5, a 1% increase in the citation count to follow-on research is associated with a small (0.03%) but statistically significant increase in the probability of the patent being a breakthrough invention.

Table 9: Patent Quality and Patent Citations to Follow-On Research

	KPSS Private Value ln(Value)		Scope Narrowness ln(Word Count)		KPST Textual Novelty Pr(Breakthrough)	
	OLS (1)	OLS (2)	OLS (3)	OLS (4)	OLS (5)	OLS (6)
ln(NPL to Follow-On)	0.046***	0.033**	-0.103**	-0.089*	0.027***	0.026***
	(0.016)	(0.016)	(0.051)	(0.048)	(0.009)	(0.009)
ln(NPL to Internal)		0.082***		-0.106***		0.005
		(0.020)		(0.023)		(0.014)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Grant-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	424,224	424,224	483,828	483,828	329,224	329,224
Avg. DV	1.960	1.960	4.894	4.894	0.190	0.190
Adjusted R^2	0.687	0.687	0.157	0.158	0.243	0.243

Notes: This table presents estimation results for the relationship between three measures of patent quality and the NPL citations to scientific publications. KPSS patent values (Kogan et al., 2017) are estimates of the private real dollar value derived from market response to patent grants. Word counts of first claims are an inverse measure of patent scope (Kuhn & Thompson, 2019). KPST breakthrough innovations are indicator variables that represent the top 10% of patents in terms of textual novelty (Kelly et al., 2021). NPLs are patent citations to scientific publications (Marx & Fuegi, 2022). NPLs to follow-on are citations to external publications that cite an upstream internal publication (up to three generations up). NPLs to internal publications are direct citations to scientific publications by the same firm. In columns 1-2, 59,643 patents are missing KPSS values. In columns 5-6, KPST indicators are available until 2010. Indicator variables for zero counts are included.
Clustered (Firm) standard-errors in parentheses. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

6.2 Firm-level Analysis

In addition to estimating the effects of follow-on research at the publication level, I explore relations between the variables at the firm level.[40] First, I study the relationship between realized follow-on research and firms' scientific publications, employment of scientists, and patenting. If follow-on research drives subsequent scientific investments and patenting outcomes at the publication level, these relations should also be observed in a firm-level panel analysis. The econometric specification is as follows:

$$\begin{aligned} \ln(Y)_{ft} = & \beta_1 \ln(\mathrm{FO}_{f,t-1}) + \\ & \beta_2 \ln(\text{Int. Patents Citing FO}_{f,t-1}) + \\ & \beta_3 \ln(\text{Ext. Patents Citing FO}_{f,t-1}) + \\ & \eta_f + \tau_t + \epsilon_{f,t} \end{aligned} \tag{4}$$

The dependent variables are annual counts of scientific publications, patents, and counts of scientists employed by firm f at year t. The main variable on the right-hand side is an accumulated stock of follow-on publications. The follow-on publications are aggregated by their own publication year to indicate realized follow-on research up to year t. The regressions also include stocks of realized patents by the focal firm and others that cite the follow-on research. I use the standard 15% depreciation value to reduce stocks over time. All models include firm and year fixed effects. Standard-errors are clustered at the firm level.[41]

Table 10 provides the estimation results. Based on a Poisson pseudo-maximum likelihood estimation (PPML), I find a positive and statistically significant relationship between realized follow-on research and subsequent scientific publications, employment of AMWS-listed scientists, award-winning scientists, and annual patenting (columns 1, 3, 5, and 7). Next, I split the stock of follow-on based on whether the firms' patents cite it. I find a more substantial relationship between the stock of used follow-on and subsequent scientific publishing, hiring, and patenting

[40] Unfortunately, the instrumental variable that I introduce at the publication level is unsuitable for aggregation up to the firm level. Therefore, I present correlational evidence in this section and acknowledge potential bias due to unobserved variables.

[41] Descriptive statistics are provided in Table B3.

compared to unused follow-on (columns 2, 4, 6, and 8). While these results cannot be interpreted as causal, they indicate that follow-on research, and especially the portion of that research cited by the firm's patents, correlates with firms' subsequent scientific investments and invention outcomes.

Lastly, I estimate the relationship between follow-on research and firm stock market value. The econometric specification is as follows:

$$\begin{aligned}\log(\text{Tobin Q})_{ft} =& \beta_1 \ln(\text{FO}_{f,t-1}) + \beta_2 \ln(\text{Int NPL to FO}_{f,t-1}) + \beta_3 \ln(\text{Ext NPL to FO}_{f,t-1}) + \\ & \beta_4 \frac{\text{Pubs}_{f,t-1}}{\text{R\&D}_{f,t-1}} + \beta_5 \frac{\text{Pats}_{f,t-1}}{\text{R\&D}_{f,t-1}} + \beta_6 \frac{\text{R\&D}_{f,t-1}}{\text{Assets}_{f,t-1}} + \\ & \eta_f + \tau_t + \epsilon_{f,t}\end{aligned} \tag{5}$$

The dependent variable is log(Tobin's Q). The main variables on the right-hand side are log counts of future follow-on publications, accumulated as stocks with 15% depreciation rates. Note that in this specification, differently from the analysis above, aggregation is done based on the focal year of publication, to reflect an expectation of follow-on research in later years. In addition, I include log counts of the focal firms' patents citing the follow-on (Internal NPL) and others' patents citing the follow-on (External NPL). I control for 1-year lagged stocks of R&D investments over assets, patents over R&D, and publications over R&D. I use the standard 15% depreciation value to reduce stocks over time. All models include firm and year fixed effects. Standard-errors are clustered at the firm level.

Table 11 provides the estimation results. Possibly surprisingly, I find a negative relation between Tobin's Q and the prospect of follow-on research (columns 1, 3, and 5). Potentially, these results are driven by limited research capabilities and greater scientific uncertainty. The negative relation becomes weaker and statistically insignificant as I account for the internal and external use of follow-on research (columns 2, 4, 6). In line with findings by Belenzon (2012) regarding patents, the prospect of internal use of follow-on research is positively correlated with Tobin's Q. In contrast, the prospect of others' use of follow-on research is negatively correlated with Tobin's Q. These results, while suggestive, support the notion that follow-on research is beneficial to the firm to the extent that it can reabsorb it in subsequent innovation, and possibly detrimental to

Table 10: Follow-On Research and Innovation Outcomes at the Firm-Year Level

	Annual Publications		Employed AMWS Scientists: All		Employed AMWS Scientists: Award-Winning		Annual Patents	
	PPML	PPML	PPML	PPML	PPML	PPML	PPML	PPML
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\ln(\text{FO Stock})_{t-1}$	0.195***	0.253***	0.060**	0.094***	0.072*	0.105**	0.055*	0.014
	(0.044)	(0.054)	(0.025)	(0.028)	(0.040)	(0.044)	(0.031)	(0.037)
$\ln(\text{Firm's pat. stock citing FO})_{t-1}$		0.145**		0.033		0.056		0.160***
		(0.058)		(0.041)		(0.057)		(0.033)
$\ln(\text{Ext. pat. stock citing FO})_{t-1}$		-0.198***		-0.057*		-0.060		-0.087***
		(0.043)		(0.029)		(0.038)		(0.027)
Time-Varying Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	26,520	26,520	20,497	20,497	8,996	8,996	35,785	35,785
Avg. DV	14.334	14.334	10.724	10.724	3.360	3.360	31.780	31.780
Psuedo R^2	0.913	0.915	0.903	0.903	0.780	0.780	0.902	0.906

Notes: This table presents estimation results for the relationship between follow-on research, subsequent scientific investments, and patenting outcomes by the originating firms. The data consists of a firm-year panel of U.S.-based publicly-owned firms between 1980 and 2015 (Arora et al., 2020). In columns 1-2, the dependent variable is a count of scientific publications authored by the focal firm in year t. In columns 3-4, the dependent variable is a count of AMWS scientists employed by the firm at year t. In columns 5-6, the count includes only award-winning scientists among AMWS. In columns 7-8, the dependent variable is a count of patents filed by the focal firm at year t. Follow-on is the stock of external papers that cite the firms' scientific publications, aggregated by the year of publication. Firm's patents citing follow-on are stocks of patents by the focal firm that cite the follow-on research. External patents citing follow-on are stocks of patents unrelated to the focal firm that cite the follow-on research. All stocks (other than scientist employment) are depreciated using an annual 15% depreciation constant. Time-varying control variables include lagged firm assets and R&D stocks. Indicator variables for zero counts are included. Observations are automatically dropped from the complete sample due to either singletons or separation by fixed effects.
Clustered (Firm) standard-errors in parentheses. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

Table 11: Follow-On Research and Firm Value

	log(Tobin's Q)					
	OLS (1)	OLS (2)	OLS (3)	OLS (4)	OLS (5)	OLS (6)
ln(1-Gen Future FO)$_{t-1}$	-0.039***	-0.032**				
	(0.009)	(0.015)				
ln(2-Gen Future FO)$_{t-1}$			-0.034***	-0.028*		
			(0.008)	(0.014)		
ln(3-Gen Future FO)$_{t-1}$					-0.031***	-0.003
					(0.007)	(0.017)
ln(Future Int. NPL to FO)$_{t-1}$		0.056***		0.050***		0.024
		(0.019)		(0.018)		(0.015)
ln(Future Ext. NPL to FO)$_{t-1}$		-0.032**		-0.033**		-0.033**
		(0.015)		(0.015)		(0.016)
(Pubs/R&D)$_{t-1}$	0.088***	0.088***	0.088***	0.086***	0.088***	0.088***
	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)
(Pats/R&D)$_{t-1}$	0.029***	0.029***	0.029***	0.029***	0.028***	0.028***
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
(R&D/Assets)$_{t-1}$	0.052***	0.052***	0.052***	0.052***	0.052***	0.052***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	43,303	43,303	43,303	43,303	43,303	43,303
Avg. DV	0.676	0.676	0.676	0.676	0.676	0.676
Adj. R^2	0.659	0.659	0.659	0.659	0.659	0.659

Notes: This table presents estimation results for the relationship between follow-on research and firm performance. The data consist of a firm-year panel of U.S.-based publicly-owned firms between 1980 and 2015 (Arora et al., 2020) The dependent variable is the log of market-to-book ratio (Tobin's Q). Control variables include one-year lagged stocks of R$D over Assets, patents over R&D, and publications over R&D. Follow-on is the stock of external citations to focal publications, aggregated to the year of publication of the focal papers. Models differ in the levels of follow-on research that are included in the count. Columns 1 and 2 include one generation of follow-on research, while columns 3-4 and 5-6 include two and three generations, respectively. All stocks are depreciated using an annual 15% depreciation constant. Internal NPL to follow-on are future patents by the focal firm that cite the follow-on research. External NPL to follow-on are future patents unrelated to the focal firm that cite the follow-on research. Indicator variables for zero counts are included.
Clustered (Firm) standard-errors in parentheses. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

the extent that others use it.

7 Discussion and Conclusions

Firms' incentives to participate in public research have received increased interest in recent years (Arora et al., 2018; Arora et al., 2021). Early works have argued that investments in science benefit firms' R&D processes by improving their combinative capabilities (Fleming & Sorenson, 2004; Kogut & Zander, 1992) and absorptive capacity (Cohen & Levinthal, 1990; Rosenberg, 1990). The literature has also suggested that firms engage with the scientific community to benefit from resources external to the firm (Cockburn & Henderson, 1998). Nonetheless, the disclosure of findings through scientific publications is typically considered to benefit the firm through channels that are not directly related to knowledge accumulation (e.g., due to scientists' preference for publishing (Stern, 2004)) or as a cost that is associated with potential knowledge spillovers to rivals. Exceptions are arguments that suggest that corporate science can influence the direction of academic research outside the firm (Hicks, 1995). The ability to influence the scientific community became increasingly important as the innovation ecosystem experienced a division of innovative labor (Arora & Gambardella, 1994) and an increased reliance on public science (Fleming et al., 2019).

This paper provides evidence that firms can benefit from external research that builds upon their prior scientific publications. First, I observe follow-on citations to firms' publications and note that they are extensively cited by the originating firms' patents. While only 7% of firms' publications are internally cited by a patent, an additional 33% of publications are cited by external follow-on research that is eventually mentioned in a patent by the same firm. This process can take years, as follow-on research becomes available and is assimilated by the firm, but it suggests that follow-on research plays an important role in connecting firms' investments in science and patenting outcomes. I also find that firms' patents that cite follow-on research are, on average, of better quality.

Next, I implement a new instrumental variable that exploits plausibly exogenous variation in scholarly attention to corporate scientific publications. Using the instrument, I find positive effects of external follow-on research on the focal firm's subsequent scientific publishing, hiring of renowned scientists, and patenting outcomes. Rivalry conditions, firms' resources and capabilities, and the nature of the scientific area moderate these effects. Follow-on research can be valuable as input into subsequent inventions or as quality validation of internally produced science. Validation seems more important when firms face higher quality uncertainty. At the firm level, I find correlational support for the benefits of follow-on research on innovation and performance.

These findings contribute to the literature on corporate science (Arora et al., 2021; Simeth & Cincera, 2016). Given the magnitude of resources and outputs produced by the scientific community, it is evident that firms can benefit from the ability to influence academics' research agenda. This paper highlights that, by participating in public research, firms can potentially influence the future content produced by academics in privately beneficial ways. The findings also relate to the literatures on open innovation and knowledge spillovers (Chesbrough, 2003; Henkel, 2006). Support is provided to the view that the opening of internal scientific knowledge could be a strategic choice that drives down R&D costs, enhances the value of complementary assets, and allows firms to mitigate uncertainty associated with their investments in science (Alexy et al., 2013; Alexy et al., 2018). Thus, this work complements recent papers that study the channels by which firms influence the scientific community (Babina et al., 2023; Bikard et al., 2019; Sohn, 2021).

This study does have several limitations. First, it conditions on the existence of a corporate publication. In this sense, it does not explore the decision to publish and does not compare the outcomes to the counterfactual of the firm refraining from publishing. As a result, I cannot directly observe the influence on academics' work. Rather, the paper shows that if there is an influence, the firm is likely to benefit from it. Future work can focus on how corporate publications affect academics and identify their incentives to build upon such work. Second, I acknowledge that the interaction coefficients in the heterogeneity analysis can reveal patterns in the data but cannot be interpreted as causal effects. Lastly, the identification of the effect is presented at the publication level. While I provide patent- and firm-level correlational evidence, a different empirical setup would be required to show causal effects on firms' financial and overall innovative performance.

Notwithstanding these limitations, the immediate implications of this study are that corporate R&D managers should consider the potential benefits that could originate from the scientific community when making decisions regarding investments in science. These decisions should align with other related strategic choices such as geographic proximity to universities (Sohn, 2021) and the funding of academic research (Babina et al., 2023). However, it is also important to note that these benefits might take years to mature, and that they might not offset completely the risks of knowledge spillovers to rivals. Under the right conditions, influence on the scientific community can result in valuable inputs that reduce the firms' R&D costs and improve innovation outcomes.

Second, the evidence also suggests that the presence of strong academic institutions can drive firms to participate in public research. Without sufficient incentives, firms might choose to refrain from investments in science, or choose secrecy over openness. Academics that are willing to engage with corporate researchers are one of the drivers of firms' contributions to the open scientific discussion. Nonetheless, whether firms' influence on the direction of scientific inquiry is socially and scientifically desirable is a subject for future work.

References

Ahmadpoor, M., & Jones, B. F. (2017). The dual frontier: Patented inventions and prior scientific advance. *Science*, *357*(6351), 583–587. https://doi.org/10.1126/science.aam9527

Alexy, O., George, G., & Salter, A. J. (2013). Cui Bono? The Selective Revealing of Knowledge and Its Implications for Innovative Activity. *Academy of Management Review*, *38*(2), 270–291. https://doi.org/10.5465/amr.2011.0193

Alexy, O., West, J., Klapper, H., & Reitzig, M. (2018). Surrendering control to gain advantage: Reconciling openness and the resource-based view of the firm. *Strategic Management Journal*, *39*(6), 1704–1727. https://doi.org/10.1002/smj.2706

Alnuaimi, T., & George, G. (2016). Appropriability and the retrieval of knowledge after spillovers. *Strategic Management Journal*, *37*(7), 1263–1279. https://doi.org/10.1002/smj.2383

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion* (1st edition). Princeton University Press.

Arora, A., Belenzon, S., & Patacconi, A. (2018). The decline of science in corporate R&D. *Strategic Management Journal*, *39*(1), 3–32. https://doi.org/10.1002/smj.2693

Arora, A., Belenzon, S., & Sheer, L. (2020). DISCERN: Duke Innovation & SCientific Enterprises Research Network. https://doi.org/10.5281/zenodo.4320782

Arora, A., Belenzon, S., & Sheer, L. (2021). Knowledge Spillovers and Corporate Investment in Scientific Research. *American Economic Review*, *111*(3), 871–898. https://doi.org/10.1257/aer.20171742

Arora, A., & Gambardella, A. (1994). The changing technology of technological change: General and abstract knowledge and the division of innovative labour. *Research policy*, *23*(5), 523–532. https://doi.org/10.1016/0048-7333(94)01003-X

Arrow, K. (1962). Economic welfare and the allocation of resources for invention. The rate and direction of inventive activity: Economic and social factors. *National Bureau of Economic Research*. https://doi.org/10.1007/978-1-349-15486-9_13

Arthur, W. B. (2011). *The Nature of Technology: What It Is and How It Evolves* (Reprint edition). Free Press.

Azoulay, P., Stuart, T., & Wang, Y. (2014). Matthew: Effect or Fable? *Management Science*, *60*(1), 92–109. https://doi.org/10.1287/mnsc.2013.1755

Babina, T., He, A. X., Howell, S. T., Perlman, E. R., & Staudt, J. (2023). Cutting the Innovation Engine: How Federal Funding Shocks Affect University Patenting, Entrepreneurship, and Publications. *The Quarterly Journal of Economics*. https://doi.org/10.1093/qje/qjac046

Baker, S., & Mezzetti, C. (2005). Disclosure as a Strategy in the Patent Race. *The Journal of Law and Economics*, *48*(1), 173–194. https://doi.org/10.1086/426879

Baruffaldi, S., & Poege, F. (2022, July 14). Like stars: How firms learn at scientific conferences. https://doi.org/10.2139/ssrn.3644106

Baruffaldi, S., Simeth, M., & Wehrheim, D. (2023). Asymmetric information and r&d disclosure: Evidence from scientific publications. *Management Science*. https://doi.org/10.1287/mnsc.2023.4721

Bednorz, J. G., & Müller, K. A. (1986). Possible highTc superconductivity in the BaLaCuO system. *Zeitschrift für Physik B Condensed Matter*, *64*(2), 189–193. https://doi.org/10.1007/BF01303701

Belenzon, S. (2012). Cumulative Innovation and Market Value: Evidence from Patent Citations. *The Economic Journal*, *122*(559), 265–285. https://doi.org/10.1111/j.1468-0297.2011.02470.x

Bikard, M., Vakili, K., & Teodoridis, F. (2019–April). When Collaboration Bridges Institutions: The Impact of University-Industry Collaboration on Academic Productivity. *Organization Science*, *30*(2), 426–445. https://doi.org/10.1287/orsc.2018.1235

Bikard, M., & Marx, M. (2020). Bridging Academia and Industry: How Geographic Hubs Connect University Science and Corporate Technology. *Management Science*, *66*(8), 3425–3443. https://doi.org/10.1287/mnsc.2019.3385

Bloom, N., Schankerman, M., & Reenen, J. V. (2013). Identifying Technology Spillovers and Product Market Rivalry. *Econometrica*, *81*(4), 1347–1393. https://doi.org/10.3982/ECTA9466

Chai, S., & Menon, A. (2019). Breakthrough recognition: Bias against novelty and competition for attention. *Research Policy*, *48*(3), 733–747. https://doi.org/10.1016/j.respol.2018.11.006

Chesbrough, H. W. (2003). *Open Innovation: The New Imperative for Creating and Profiting from Technology*. Harvard Business Press.

Cockburn, I. M., & Henderson, R. M. (1998). Absorptive Capacity, Coauthoring Behavior, and the Organization of Research in Drug Discovery. *The Journal of Industrial Economics*, *46*(2), 157–182. https://doi.org/10.1111/1467-6451.00067

Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. *Administrative Science Quarterly*, 128–152. https://doi.org/10.2307/2393553

Cohen, W. M., Nelson, R. R., & Walsh, J. P. (2002). Links and Impacts: The Influence of Public Research on Industrial R&D. *Management Science*, *48*(1), 1–23. https://doi.org/10.1287/mnsc.48.1.1.14273

Cohn, J. B., Liu, Z., & Wardlaw, M. I. (2022). Count (and count-like) data in finance. *Journal of Financial Economics*, *146*(2), 529–551. https://doi.org/10.1016/j.jfineco.2022.08.004

Cornelissen, T., Dustmann, C., Raute, A., & Schönberg, U. (2016). From LATE to MTE: Alternative methods for the evaluation of policy interventions. *Labour Economics*, *41*, 47–60. https://doi.org/10.1016/j.labeco.2016.06.004

Correia, S., Guimarães, P., & Zylkin, T. (2020). Fast Poisson estimation with high-dimensional fixed effects. *The Stata Journal*, *20*(1), 95–115. https://doi.org/10.1177/1536867X20909691

Dahlander, L., & Wallin, M. W. (2006). A man on the inside: Unlocking communities as complementary assets. *Research Policy*, *35*(8), 1243–1259. https://doi.org/10.1016/j.respol.2006.09.011

Dasgupta, P., & David, P. A. (1994). Toward a new economics of science. *Research Policy*, *23*(5), 487–521. https://doi.org/10.1016/0048-7333(94)01002-1

David, P. A. (1998). Common Agency Contracting and the Emergence of "Open Science" Institutions. *The American Economic Review*, *88*(2), 15–21.

Erdelez, S. (1999). Information Encountering: It's More Than Just Bumping into Information. *Bulletin of the American Society for Information Science and Technology*, *25*(3), 26–29. https://doi.org/10.1002/bult.118

Fleming, L., Greene, H., Li, G., Marx, M., & Yao, D. (2019). Government-funded research increasingly fuels innovation. *Science*, *364*(6446), 1139–1141. https://doi.org/10.1126/science.aaw2373

Fleming, L., & Sorenson, O. (2004). Science as a map in technological search. *Strategic Management Journal*, *25*(8-9), 909–928. https://doi.org/10.1002/smj.384

Foster, A., & Ford, N. (2003). Serendipity and information seeking: An empirical study. *Journal of Documentation*, *59*(3), 321. https://doi.org/10.1108/00220410310472518

Gavetti, G., Helfat, C. E., & Marengo, L. (2017). Searching, shaping, and the quest for superior performance. *Strategy Science*, *2*(3), 194–209. https://doi.org/10.1287/stsc.2017.0036

Harhoff, D., Henkel, J., & von Hippel, E. (2003). Profiting from voluntary information spillovers: How users benefit by freely revealing their innovations. *Research Policy*, *32*(10), 1753–1769. https://doi.org/10.1016/S0048-7333(03)00061-1

Helfat, C. E., Kaul, A., Ketchen Jr., D. J., Barney, J. B., Chatain, O., & Singh, H. (2023). Renewing the resource-based view: New contexts, new concepts, and new methods. *Strategic Management Journal*, *44*(6), 1357–1390. https://doi.org/10.1002/smj.3500

Henkel, J. (2006). Selective revealing in open innovation processes: The case of embedded Linux. *Research Policy*, *35*(7), 953–969. https://doi.org/10.1016/j.respol.2006.04.010

Hicks, D. (1995). Published papers, tacit competencies and corporate management of the public/private character of knowledge. *Industrial and corporate change*, *4*(2), 401–424. https://doi.org/10.1093/icc/4.2.401

Hudson, J. (2007). Be known by the company you keep: Citations — quality or chance? *Scientometrics*, *71*(2), 231–238. https://doi.org/10.1007/s11192-007-1671-6

Jaffe, A. B. (1989). Real Effects of Academic Research. *The American Economic Review*, *79*(5), 957–970.

Jee, S. J., & Sohn, S. Y. (2023). A firm's creation of proprietary knowledge linked to the knowledge spilled over from its research publications: The case of artificial intelligence. *Industrial and Corporate Change*, dtad002. https://doi.org/10.1093/icc/dtad002

Jin, C., Ma, Y., & Uzzi, B. (2021). Scientific prizes and the extraordinary growth of scientific topics. *Nature Communications*, *12*(1), 5619. https://doi.org/10.1038/s41467-021-25712-2

Kelly, B., Papanikolaou, D., Seru, A., & Taddy, M. (2021). Measuring Technological Innovation over the Long Run. *American Economic Review: Insights*, *3*(3), 303–320. https://doi.org/10.1257/aeri.20190499

Kogan, L., Papanikolaou, D., Seru, A., & Stoffman, N. (2017). Technological Innovation, Resource Allocation, and Growth. *The Quarterly Journal of Economics*, *132*(2), 665–712. https://doi.org/10.1093/qje/qjw040

Kogut, B., & Zander, U. (1992). Knowledge of the firm, combinative capabilities, and the replication of technology. *Organization science*, *3*(3), 383–397. https://doi.org/10.1287/orsc.3.3.383

Kuhn, J. M., & Thompson, N. C. (2019). How to Measure and Draw Causal Inferences with Patent Scope. *International Journal of the Economics of Business*, *26*(1), 5–38. https://doi.org/10.1080/13571516.2018.1553284

Leiponen, A., & Helfat, C. E. (2010). Innovation objectives, knowledge sources, and the benefits of breadth. *Strategic Management Journal*, *31*(2), 224–236. https://doi.org/10.1002/smj.807

Lund, B. D., & Maurya, S. K. (2020). The relationship between highly-cited papers and the frequency of citations to other papers within-issue among three top information science journals. *Scientometrics*, *125*(3), 2491–2504. https://doi.org/10.1007/s11192-020-03720-1

Makri, S., & Blandford, A. (2012). Coming Across Information Serendipitously - Part 1: A Process Model. *Journal of Documentation*, *68*. https://doi.org/10.1108/00220411211256030

Mansfield, E. (1991). Academic research and industrial innovation. *Research Policy*. https://doi.org/10.1016/0048-7333(91)90080-a

Marx, M., & Fuegi, A. (2022). Reliance on Science in Patenting. https://doi.org/10.5281/zenodo.6468792

Nelson, R. R. (1959). The simple economics of basic scientific research. *Journal of political economy*, *67*(3), 297–306.

Nelson, R. R., & Winter, S. G. (1982). The Schumpeterian Tradeoff Revisited. *The American Economic Review*, *72*(1), 114–132.

Ocasio, W. (1997). Towards an Attention-Based View of the Firm. *Strategic Management Journal*, *18*(S1), 187–206. https://doi.org/10.1002/(SICI)1097-0266(199707)18:1+⟨187::AID-SMJ936⟩3.0.CO;2-K

Perkmann, M., Salandra, R., Tartari, V., McKelvey, M., & Hughes, A. (2021). Academic engagement: A review of the literature 2011-2019. *Research Policy*, *50*(1), 104114. https://doi.org/10.1016/j.respol.2020.104114

Perkmann, M., Tartari, V., McKelvey, M., Autio, E., Broströöm, A., D'Este, P., Fini, R., Geuna, A., Grimaldi, R., Hughes, A., Krabel, S., Kitson, M., Llerena, P., Lissoni, F., Salter, A., & Sobrero, M. (2013). Academic engagement and commercialisation: A review of the literature on university–industry relations. *Research Policy*, *42*(2), 423–442. https://doi.org/10.1016/j.respol.2012.09.007

Polidoro, F., & Theeke, M. (2012). Getting Competition Down to a Science: The Effects of Technological Competition on Firms' Scientific Publications. *Organization Science*, *23*(4), 1135–1153. https://doi.org/10.1287/orsc.1110.0684

Prato, M., & Ferraro, F. (2018). Starstruck: How Hiring High-Status Employees Affects Incumbents' Performance. *Organization Science*, *29*(5), 755–774. https://doi.org/10.1287/orsc.2018.1204

Roach, M., & Cohen, W. M. (2012). Lens or Prism? Patent Citations as a Measure of Knowledge Flows from Public Research. *Management Science*, *59*(2), 504–525. https://doi.org/10.1287/mnsc.1120.1644

Rosenberg, N. (1990). Why do firms do basic research (with their own money)? *Research Policy*, *19*(2), 165–174. https://doi.org/10.1016/0048-7333(90)90046-9

Rotolo, D., Camerani, R., Grassano, N., & Martin, B. R. (2022). Why do firms publish? A systematic literature review and a conceptual framework. *Research Policy*, *51*(10), 104606. https://doi.org/10.1016/j.respol.2022.104606

Shen, X. (2022). Spread The Seed to Win a Long Harvest. *Academy of Management Proceedings*, *2022*(1), 14751. https://doi.org/10.5465/AMBPP.2022.14751abstract

Simcoe, T. S., & Waguespack, D. M. (2011). Status, Quality, and Attention: What's in a (Missing) Name? *Management Science*, *57*(2), 274–290. https://doi.org/10.1287/mnsc.1100.1270

Simeth, M., & Cincera, M. (2016). Corporate Science, Innovation, and Firm Value. *Management Science*, *62*(7), 1970–1981. https://doi.org/10.1287/mnsc.2015.2220

Simeth, M., & Raffo, J. D. (2013). What makes companies pursue an Open Science strategy? *Research Policy*, *42*(9), 1531–1543. https://doi.org/10.1016/j.respol.2013.05.007

Sinha, A., Shen, Z., Song, Y., Ma, H., Eide, D., Hsu, B.-J. (, & Wang, K. (2015). An Overview of Microsoft Academic Service (MAS) and Applications. *Proceedings of the 24th International Conference on World Wide Web*, 243–246. https://doi.org/10.1145/2740908.2742839

Sohn, E. (2021). How Local Industry R&D Shapes Academic Research: Evidence from the Agricultural Biotechnology Revolution. *Organization Science*, *32*(3), 675–707. https://doi.org/10.1287/orsc.2020.1407

Stern, S. (2004). Do scientists pay to be scientists? *Management science*, *50*(6), 835–853. https://doi.org/10.1287/mnsc.1040.0241

Wooldridge, J. M. (2010, October 1). *Econometric analysis of cross section and panel data, second edition* (second edition). The MIT Press.

Yang, H., Phelps, C., & Steensma, H. K. (2010). Learning from What Others Have Learned from You: The Effects of Knowledge Spillovers on Originating Firms. *Academy of Management Journal*, *53*(2), 371–389. https://doi.org/10.5465/amj.2010.49389018

Yang, H., & Steensma, H. K. (2014). When do firms rely on their knowledge spillover recipients for guidance in exploring unfamiliar knowledge? *Research Policy*, *43*(9), 1496–1507. https://doi.org/10.1016/j.respol.2014.04.016

Appendices

Appendix A Case Studies

A.1 Overview

This section provides multiple examples of follow-on research and its value for the originating firms. These examples were obtained by closely following scientific and patent citations within the data and combining them with complementary details from online web searches and correspondence with the authors.

In the first example, follow-on research by academics was used as an input into subsequent innovation by the originating firm. The firm engaged with the scientific community by regularly publishing scientific findings. Consequentially, external findings that cited the firms' publications were further developed by the firm and incorporated in it's patents. The example describes an incremental advancement to organic light-emitting diode (OLED) technology that was developed at the Hong Kong University of Science and Technology and subsequently implemented in a patent by Universal Display Company (UDC).

In the second example, follow-on research by academics lead to a collaboration between them and the originating firm. The example describes the development of Morpholino Oligomers, a type of molecular structure that can bind to genetic material. These structures were initially developed by AVI BioPharma (subsequently Sarepta Therapeutics), a private biotech company based in Cambridge, Massachusetts. Following this development, a scientific research group at the University of Western Australia found a therapeutic opportunity for Morpholinos in treating Duchenne, a type of muscular dystrophy. The findings were eventually licensed, further developed, and commercialized by Sarepta.

The third example is a case from software development. Here, the focal firm described an application that served as a motivation for academics to develop relevant upstream algorithms. Such external developments then served as a baseline in one of the firm's own patents to test the performance of their technology and compare it to technologies that were available to others. The example describes the development of algorithms for video object segmentation (breaking up video footage into objects) by Adobe Inc. Later, a patent by the firm used several external algorithms to establish the case for better performance of their own technology.

The fourth example is a case from nanotechnology. In this example, the focal firm made a scientific discovery and intended to use it in specific applications. Academics at other institutions produced follow-on findings that were applicable to a completely different set of products. As a result, various assignees filed patents that cited the follow-on findings. However, while the original discovery was patented by the focal firm, it seems not to have used the follow-on findings in subsequent innovation. Possibly, the firm lacked interest in other product lines that were unrelated to its own, or lacked the complementarities needed to benefit from them. The example describes the development of a technology that allows the tuning of liquid on nanostructured surfaces by a research group at Bell Labs.

In the fifth example, a firm's publications were hardly cited by external researchers, and the follow-on research that did emerge was not mentioned by the originating firm's patents. Possibly, in this case the firm's incentives to engage with the scientific community were not related to the potential usefulness of follow-on research.

A.2 Example from Electronics: Silver Film Improves Color Saturation in OLEDs

Universal Display Corporation was founded in 1994 by Sherwin Seligsohn with the goal of developing displays that are based on organic light-emitting diodes (OLED).[42] Benefiting from long-term research

[42] About UDC. *Universal Display Corporation.* Retrieved July 14, 2022, from https://oled.com/about/

contracts with Princeton University, the company became a leader in OLED technology. Today, almost all OLED products incorporate proprietary technologies owned by UDC. Throughout the years, UDC focused on advancing OLED technologies through investments in scientific research and collaborations with other organizations.

Researchers at UDC frequently publish the firm's scientific findings. The example described here involves four of the firm's papers, published between the years 2000 and 2002: Burrows, Forrest, et al. (2000) found that long lifetimes are an intrinsic property of phosphorescent OLEDs; Chwang et al. (2002) studied the performance of graded mixed-layer OLEDs, a design that the authors suggested could extend the device's lifetime and make them applicable to flat panel displays; Lu et al. (2002) studied top-emitting OLEDs, a design that was proved to be 20% more efficient than equivalent bottom-emitting OLEDs; Lastly, Burrows, Gu, et al. (2000) studied semitransparent cathodes in OLEDs that can be applied to various use-cases.

Common to all publications above, is that they were cited directly or indirectly by Peng et al. (2005), which was authored by researchers at the Hong Kong University of Science and Technology. The citations suggest that the researchers built on UDC's prior findings in their work. In addition, Peng et al. (2005) and three of the publications above were published in Applied Physics Letters (the other two were published in the Journal of Applied Physics). This relation further supports the notion that researchers at UDC and at the Hong Kong University of Science and Technology were part of the same scientific community.

In their paper, Peng et al. (2005) studied several metals as alternatives to the use of indium-tin oxide (ITO), the material that was typically used before as the anode material. They found that silver can serve as an effective alternative to ITO. In their experiments, silver anode resulted in improved current voltage and optical performance. They suggested, theoretically, that the use of semitransparent silver anodes can enhance light extraction efficiency. However, the authors acknowledged that further investigation is necessary.

In 2007, two years after the publication of Peng et al. (2005), a researcher at UDC filed for a patent that built on their proposal to use semitransparent silver in OLEDs.[43] The patent suggested to use silver (or other relevant metals) as a color saturation enhancement layer between the two electrodes of the OLED device:

> It is believed that certain metals, such as aluminum and chromium, are generally thought of as undesirable for placement in organic light emitting devices between the anode and the organic layers. See, Peng et al., Efficient organic light emitting diode using semitransparent silver as anode, Applied Physics Letters 87, 173505, p. 1 (2005), teaching that high work function metals are desirable to lower barriers for hole injection; conversely, low work function metals are not desirable. Surprisingly, it has been found that these metals may be used as thin layers between the anode and the organic layers of an organic light emitting device as color Saturation enhancement layers... Silver is also a preferred material for use as a color Saturation enhancement layer.

To summarize, in this example, an academic research group was influenced by the scientific publications originating from UDC. Then, their work (Peng et al., 2005) served as an input into subsequent innovation by the firm.[44]

---

[43] D'Andrade, B. (2013). Saturated color organic light emitting devices (United States Patent No. US8476822B2).

[44] The details in this example were verified with the corresponding author of Peng et al. (2005).

A.3 Example from Bio-pharmaceuticals: Morpholino Oligomers and the Treatment of Duchenne Disease

During the 1970s several research groups started working on antisense therapeutics strategies for binding to genetic material (Summerton, 2016). These strategies, it was suggested, could offer treatment to a wide range of conditions, including viral diseases, cancers and genetic defects. The first patent in this area was filed in 1977 by Summerton and Bartlett.[45] In 1980, Summerton founded Antivirals Incorporated (later AVI BioPharma and subsequently Sarepta Therapeutics), the first company that focused on developing and commercializing these treatment methods. In 1985, with advice from Dwight Weller, Summerton developed a molecule structure that radically departed from previous designs. This design, named *morpholino oligomers*, used far cheaper materials and was easier to produce in comparison to previously existing designs. Between 1985 and the mid 1990s, AVI BioPharma and other research groups further developed and enhanced the morpholino molecular structure. In studies on cultured human cells, morpholinos outperformed the competition in both efficacy and specificity.

Once the promise of morpholino oligomers has been sufficiently established, Summerton and Weller published a review article that covered details on the design, preparation and properties of these structures (Summerton & Weller, 1997). They also started exploring therapeutic applications, such as increasing hemoglobin production in blood cells of thalassemic patients (Lacerra et al., 2000).[46]

In the early 2000s, The development of morpholino oligmers was picked up by a research group at the Centre for Neuromuscular and Neurological Disorders in the University of Western Australia, who were working on developing an antisense-based therapy to Duchenne muscular dystrophy (DMD). The group noted:

> One chemistry that is gaining wide recognition for use in antisense applications is the morpholino oligonucleotide developed by Summerton and Weller. These authors developed the morpholino structural type with the intention that this chemistry could provide several advantages in the clinical application(s) of antisense therapeutics, such as strong nucleic acid binding, resistance to nucleases, minimal nonantisense effects, high aqueous solubility and relatively low synthesis costs (Gebski et al., 2003).

However, an important challenge that faced the researchers was the delivery of the morpholinos into the cell nucleus. To overcome this challenge, the researchers at the University of Western Australia suggested to anneal the morpholinos with additional DNA/RNA molecules, or 'leashes'. Along with filing patents to protect their inventions, they described their findings in scientific publications:[47]

> The uncharged backbone compromises delivery, for non-ionic AOs cannot easily be delivered into cultured cells using delivery agents such as cationic liposomes. To circumvent this difficulty, we investigated the use of single stranded (anionic) nucleic acid 'leashes' which were annealed to the morpholino AO, allowing the AO : leash duplex to be complexed with Lipofectin. (Gebski et al., 2003).

Following the findings of Gebski et al. (2003), Sarepta further developed methods for treating Duchenne using morpholino oligomers and studied additional applications to other related diseases. In 2011, Sarepta filed a patent for the treatment of myotonic dystrophy, another type of muscular dystrophy along with Duchenne:

[45] Summerton, J. E., & Bartlett, P. A. (1978). Nucleic acid crosslinking agent and affinity inactivation of nucleic acids therewith (United States Patent No. US4123610A).

[46] Thalassemia is an inherited blood disorder that causes hemoglobin deficiencies.

[47] Fletcher, S., McClorey, G., & Wilton, S. (2004). *Antisense Oligonucleotides for Inducing Exon Skipping and Methods of Use Thereof* (AU2004903474A0).

> The parent application disclosed and claimed the use of these two CPPs[48] for targeting antisense oligonucleotides to muscle tissue, in treating certain muscle pathologies. For example, in treating Duchenne muscular dystrophy (DMD) ... The present invention applies this strategy additionally to the treatment of myotonic dystrophy MD1 and MD2 in muscle tissue, including skeletal and heart muscle tissue.[49]

In this patent, the inventors acknowledge the complementarities between the focal invention and the prior scientific findings by the research group at the University of Western Australia:

> The oligonucleotide-(RXRR(B/X)R)$_2$XB conjugate compounds of the invention may be used in conjunction with homing peptides selective for the target tissue, to further enhance muscle-specific delivery. An example of this approach can be found in the application of muscle-binding peptides (Samoylova and Smith, 1999; Vodyanoy et al., U.S. Appn. Pubn. No. 2003064.0466) coupled to antisense oligomers designed to be therapeutic treatments for Duchenne muscular dystrophy (DMD) (Gebski, Mann et al. 2003; Alter, Lou et al. 2006) (PCT Pubn. No. WO2006000057).

In 2013, the complementarities between the inventions at Sarepta and the Duchenne treatments developed at the University of Western Australia eventually lead to an exclusive licensing agreement for the commercialization of the treatment:[50]

> Sarepta has an exclusive, worldwide licensing agreement with the University of Western Australia (UWA) for intellectual property rights to support the development of exon-skipping drug candidates for the treatment of Duchenne muscular dystrophy (DMD). The agreement grants Sarepta rights to UWA's extensive patent portfolio in DMD and enables the Company to expand its exon-skipping pipeline with new candidates to address the majority of patients with DMD worldwide. [51]

At last, in 2016, the U.S. Food and Drug Administration (FDA) approved Sarepta's product as the first drug for treating patients with Duchenne:

> The U.S. Food and Drug Administration today approved Exondys 51 (eteplirsen) injection, the first drug approved to treat patients with Duchenne muscular dystrophy (DMD). Exondys 51 is specifically indicated for patients who have a confirmed mutation of the dystrophin gene amenable to exon 51 skipping, which affects about 13 percent of the population with DMD.[52]

---

[48] Cell penetrating peptides

[49] Moulton, H. M., & Kole, R. (2014). *Compound and method for treating myotonic dystrophy* (United States Patent No. US8741863B2).

[50] Sarepta Therapeutics and University of Western Australia Announce Exclusive Worldwide Licensing Agreement for Exon-Skipping Program in Duchenne Muscular Dystrophy. Retrieved July 14, 2022, from https://investorrelations.sarepta.com/news-releases/news-release-details/sarepta-therapeutics-and-university-western-australia-announce

[51] Sarepta Strategic Partnerships. Retrieved July 14, 2022, from https://www.sarepta.com/science/strategic-partners

[52] FDA grants accelerated approval to first drug for Duchenne muscular dystrophy. FDA; Retrieved July 14, 2022, from https://www.fda.gov/news-events/press-announcements/fda-grants-accelerated-approval-first-drug-duchenne-muscular-dystrophy

Today, Duchenne is one of the core disease areas in Sarepta's portfolio and ongoing research is conducted to further develop treatments for additional Duchenne subtypes as well as other related diseases. In 2021, Sarepta's total revenue from their portfolio of treatments and collaborations surpassed $700 million.[53]

A.4 Example from Software Development: Image Segmentation Algorithms

Adobe Inc. is a computer software company that was founded in 1982 in California. It is a leader in specialized software for a wide range of creative content creation. up until the 2010s, the company focused on developing tools for professional artists, designers and video editors. The introduction of cameras on mobile phones, advancements in computer processing power and the rise of online sharing platforms (e.g. Youtube) enabled the general public to take part in content creation, and demand for simple and intuitive editing tools increased. Goldman et al. (2008) is a scientific publication by Adobe developers that suggested that Adobe was interested at the time in developing tools for video annotation and composition that are more user-friendly. However, the development of these tools required a complex preprocessing step that is known as video object segmentation. This process allows the software to automatically identify objects in the video and separate them from other objects and the background:

> In this paper we propose a framework using (2D) video object motion to enable novel approaches to user interaction... In particular, we propose novel interfaces for three tasks that are under-served by present-day video interfaces: annotation, navigation, and image composition.
>
> ... To achieve these interactions, our system first analyzes the video in a fully automatic preprocessing step that tracks the motion of image points across the video and segments those tracks into coherently moving groups.

Brendel and Todorovic (2009), a publication by researchers at Oregon State University, cited Goldman et al. (2008) as a motivation for developing better video object segmentation methods. The authors claimed that the method used by Adobe suffered from several drawbacks and offered an improved solution:

> This paper presents an approach to unsupervised video object segmentation (VOS). Our goal is to delineate the boundaries of all moving and static objects occurring in an arbitrary video... VOS is a prerequisite step of a wide range of higher level vision algorithms, including activity recognition video summarization and retrieval, and nonphotorealistic video rendering.
>
> ...
>
> Currently, the two predominant approaches to VOS are tracking interest points, and perceptual grouping of pixels from all frames. There is a number of unsatisfying aspects about both of them. Point-based approaches group the trajectories of keypoints with similar motions. However, tracking points yields only a confidence map of the objects' vicinity – not segmentation.
>
> ...
>
> In this paper, we adopt an alternative, hybrid formulation.

[53] Sarepta Therapeutics Announces Fourth Quarter and Full-Year 2021 Financial Results and Recent Corporate Developments. Retrieved July 14, 2022, from https://investorrelations.sarepta.com/news-releases/news-release-details/sarepta-therapeutics-announces-fourth-quarter-and-full-year-2021

Next, Grundmann et al. (2010), a collaboration of researchers at Georgia Institute of Technology and Google Research, identified three main challenges in video object segmentation: temporal coherence, automatic processing and scalability. They adopted a method of image segmentation and generalized it for video processing. Importantly, they claimed that their method outperforms previous methods, including the one developed by Brendel and Todorovic (2009):

> Tracking-based video segmentation methods generally define segments at frame-level and use motion, color and spatial cues to force temporal coherence. Following the same line of work, Brendel and Todorovic (2009) used contour cues to allow splitting and merging of segments to boost the tracking performance.
>
> ...
>
> Our novel video segmentation algorithm addresses all of the above challenges. We build a 3-D graph from the video volume and generalize Felzenszwalb and Huttenlocher's graph-based image segmentation to obtain an initial oversegmentation of the video volume into relatively small space-time regions.

Joulin et al. (2012) offered methods for cosegmentation, an additional development related to the methods discussed above, in which availability of multiple images (such as in a video sequence) offers a type of supervision that can improve the segmentation process. They found that their methods performed well with the dataset provided by Grundmann et al. (2010):

> The aim of cosegmentation methods is to simultaneously divide a set of images assumed to contain instances of K different object classes into regions corresponding to these classes. Note that in this context, an "object" may refer to what is usually called a "thing" (a car, a cow, etc.) but might also be a texture (grass, rocks), or other "stuff" (a building, a forest)... The proposed approach has been implemented and tested on several datasets including video sequences.
>
> ...
>
> Our experiments with iCoseg suggest that our method is particularly well suited to keyframes from the same video shot, since these are likely to feature the same objects under similar illumination. This is confirmed with our experiments with two short video clips taken from the Hollywood-2 and Grundmann datasets.

The last part of this example is Cohen, Price, and Ahmed (2015), a patent filed in 2013 by researchers at Adobe Inc. The patent claimed priority on a method in which a user provides an input (such as a "dog") and the system automatically segments the corresponding object from within the video:

> Techniques are disclosed herein that enable digital images to be segmented based on a user's semantic input. In other words, given an input image of a person walking a dog adjacent to a tree, a user can simply provide the semantic input "dog" and the system will segment the dog from the other elements in the image. If the user provides other semantic input, such as "person" or "tree", the system will segment the person or the tree, respectively, from the same image.

The inventors used prior developments in object segmentation, including the methods developed by Joulin et al. (2012), as benchmarks for their own approach:

> The Jaccard similarity coefficient J_s corresponding to segmentation using an example embodiment disclosed herein was compared with a corresponding coefficient resulting from segmentation using four different cosegmentation techniques...The compared cosegmentation algorithms are described in: Joulin et al., "Multi-Class Cosegmentation". Proceedings of IEEE ComputerVision and Pattern Recognition (CVPR 2012), pp. 542-549 (2012) ("Joulin-1");...
>
> The results of the foregoing comparison are listed in Table A. In particular, Table A illustrates that the tested example embodiment provides a segmentation that is significantly more accurate than the compared cosegmentation techniques in most applications.

To summarize, in 2008 Adobe developed tools to meet the demand for user-friendly and intuitive video editing. These tools required to solve a complex problem of video segmentation, and the methods available then suffered from various issues. Research groups outside of Adobe were aware of the demand for better methods and developed various solutions to these problems. Eventually, Adobe patented additional techniques and used external solutions as a baseline for comparison.

A.5 Example from Nanotechnology: Dynamic Tuning of Liquids on Nanostructured Surfaces

In 2006, the world-renowned Bell Laboratories were split from AT&T and were placed under a new company named Lucent Technologies. Under Lucent, researchers at Bell Labs continued to conduct scientific research in a wide array of areas. In 2004, a research group led by Prof. Krupenkin at Bell Labs discovered a electrical method to dynamically control the behaviour of liquids on nanostructured surfaces. The original publication by Krupenkin et al. (2004) was accompanied by a paired patent (Kornblit et al., 2016). In this study, the research group noted the wide range of potential applications of these findings:

> In this work, we propose a new approach that allows us to achieve effective electrowetting on nanostructured superhydrophobic surfaces...The ability to dynamically change the interaction between the liquid and the nanostructured substrate potentially opens a wide range of exciting new applications. The particular areas of interest include microfluidics, lab-on-a-chip devices, chemical microreactors, thermal management of microelectronics, drag reduction systems, and optical communications, as well as many others.

The work on tunable nanostructured surfaces won Prof. Krupenkin the American Chemical Society Industrial Innovation Award in 2007.[54] While the invention could be potentially applied to multiple products, it seems that at that time Lucent envisioned its use in developing technologies to enhance power cell batteries. In a press release, Bell Labs announced a partnership with mPhase Technologies to use the discovery for battery development:

> Bell Labs scientists and engineers recently made a significant breakthrough in microfluidics that enables dynamic control of surfaces when interacting with a liquid - a key enabler for making "Smart Batteries" a reality. Fine control of liquids at the micro and macro scale will allow scientists to create batteries that can be activated upon demand.[55]

[54] Krupenkin Group. Retrieved on August 11, 2022 from http://www.krupenkin.com/people/people.aspx

[55] MPhase and Bell Labs to Develop Nanotech Power Cell Batteries - New Technology. AZoNano.com, 22 Mar. 2004. Retrieved July 14, 2022, from https://www.azonano.com/article.aspx?ArticleID=654.

The same research team at Lucent also filed for patents that use the original discovery in new battery technologies (Hodes et al., 2010):

> A battery having an electrode with at least one nanostructured Surface is disclosed wherein the nanostructured Surface is divided into cells and is disposed in a way Such that an electrolyte fluid of the battery is prevented from contacting the portion of electrode associated with each cell. When a Voltage is passed over the nanostructured Surface associated with a particular cell, the electrolyte fluid is caused to penetrate the nanostructured surface of that cell and to contact the electrode, thus activating the portion of the battery associated with that cell.

Meanwhile, the original paper by Krupenkin et al. (2004) was cited externally 274 times since its publication. These papers were later cited 14,077 times. In the third generation of citations there are over 180 thousand citations. While these publications were cited by over 2,400 unique patents, none of these patents are assigned to Bell Labs or Lucent. This is an indication that follow-on research was not used by Bell Labs in related subsequent innovation.

It is hard to know the exact reason why Bell Labs did not use the follow-on research originating from their own discovery. However, a close examination of the topics in follow-on publications could reveal a possible explanation – the subsequent technologies were unrelated to their lines of business. Numerous studies that cite Krupenkin et al. (2004) focused on developing techniques to enhance "lab-on-a-chip" applications. For example, The Wheeler Microfluidics Laboratory, a research group at Toronto University, published several follow-on papers to Krupenkin et al. (2004). In one such paper, Luk et al. (2008) suggested that Pluronics additives can solve stickiness issues in digital microfluidics. The publication was accompanied by several patents assigned to the same research group (e.g. Wheeler and Jebrail (2011)). These findings and inventions might not have been of interest for researchers and companies that were focused on battery-enhancing technologies.

This example suggests that firms with a wide range of product lines, adaptable business models and advanced commercialization capabilities could more easily benefit from the breadth of applications that external follow-on research could provide. Firms that are focused on specific products might benefit from licensing their upstream inventions in a market for technology, but might be limited in incorporating external research in their own subsequent inventions.

A.6 Example from Radio Engineering: Dielectric Loaded Hybrid Mode Horn Antennas

Lockheed Martin is an American corporation that is a leader in developing aerospace and defense technologies. Among firms with the largest number of scientific publications, Lockheed has a very low percentage of publications that are cited by external follow-on publications and then used in the firm's own patents (only 15%, compared with 41% for Amgen Inc. and 31% for Hewlett-Packard). A close look at the scientific publications by Lockheed could provide an explanation.

For example, Lier and Kishk (2005) is a collaborative publication by a researcher at Lockheed and a researcher at the University of Mississippi. It describes a new model for a very specific type of antenna. This antenna could be used on planes, satellites and reflector antennas:

> A new class of hybrid mode horn antennas, which can be designed for a specific gain or sidelobe requirement and low cross-polarization, has been presented...It could be particularly useful in millimeter wave applications since the design is compliant with small size manufacturing. Finally, the flat top pattern design makes it a candidate earth coverage horn on-board satellites and a candidate feed for reflector antennas with enhanced directivity.

This publication was accompanied by several patents by the same author-inventor (Lier, 2009, 2012, 2013; Lier & Katz, 2009, 2011). The patents provided protection to a wide set of related inventions: a low index metamaterial, artificial dielectric antenna elements, horn antennas and antenna arrays.

While internally useful, the original publication received only 16 external scientific citations (171 across three generations). None of these publications were mentioned in any of Lockheed Martin's subsequent patents and only a handful were cited by Lockheed in subsequent scientific publications. An explanation for this is that possibly this research area is dominated by corporate researchers. Few academics at public research institutions have the relevant knowledge and interest in developing these findings outside the original firm. Therefore, it is likely that Lockheed published the focal publication without expecting to engage with useful external follow-on findings.

References

Brendel, W., & Todorovic, S. (2009). Video object segmentation by tracking regions. *2009 IEEE 12th International Conference on Computer Vision*, 833–840. https://doi.org/10.1109/ICCV.2009.5459242

Burrows, P. E., Forrest, S. R., Zhou, T. X., & Michalski, L. (2000). Operating lifetime of phosphorescent organic light emitting devices. *Applied Physics Letters*, *76*(18), 2493–2495. https://doi.org/10.1063/1.126386

Burrows, P. E., Gu, G., Forrest, S. R., Vicenzi, E. P., & Zhou, T. X. (2000). Semitransparent cathodes for organic light emitting devices. *Journal of Applied Physics*, *87*(6), 3080–3085. https://doi.org/10.1063/1.372303

Chwang, A. B., Kwong, R. C., & Brown, J. J. (2002). Graded mixed-layer organic light-emitting devices. *Applied Physics Letters*, *80*(5), 725–727. https://doi.org/10.1063/1.1446992

Cohen, S., Price, B. L., & Ahmed, E. (2015). *Semantic object proposal generation and validation* (US9129192B2).

Gebski, B. L., Mann, C. J., Fletcher, S., & Wilton, S. D. (2003). Morpholino antisense oligonucleotide induced dystrophin exon 23 skipping in mdx mouse muscle. *Human Molecular Genetics*, *12*(15), 1801–1811. https://doi.org/10.1093/hmg/ddg196

Goldman, D. B., Gonterman, C., Curless, B., Salesin, D., & Seitz, S. M. (2008). Video object annotation, navigation, and composition. *Proceedings of the 21st Annual ACM Symposium on User Interface Software and Technology - UIST '08*, 3. https://doi.org/10.1145/1449715.1449719

Grundmann, M., Kwatra, V., Han, M., & Essa, I. (2010). Efficient hierarchical graph-based video segmentation. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2141–2148. https://doi.org/10.1109/CVPR.2010.5539893

Hodes, M. S., Kolodner, P. R., Krupenkin, T. N., Lyons, A. M., Mandich, M. L., Taylor, J. A., & Weiss, D. (2010). *Reserve cell-array nanostructured battery* (US7785733B2).

Joulin, A., Bach, F., & Ponce, J. (2012). Multi-class cosegmentation. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 542–549. https://doi.org/10.1109/CVPR.2012.6247719

Kornblit, A., Krupenkin, T. N., Mandich, M. L., Schneider, T. M., Taylor, J. A., & Yang, S. (2016). *Apparatus for controlling the movement of a liquid on a nanostructured or microstructure surface* (US9433907B2).

Krupenkin, T. N., Taylor, J. A., Schneider, T. M., & Yang, S. (2004). From Rolling Ball to Complete Wetting: The Dynamic Tuning of Liquids on Nanostructured Surfaces. *Langmuir*, *20*(10), 3824–3827. https://doi.org/10.1021/la036093q

Lacerra, G., Sierakowska, H., Carestia, C., Fucharoen, S., Summerton, J., Weller, D., & Kole, R. (2000). Restoration of hemoglobin A synthesis in erythroid cells from peripheral blood of thalassemic patients. *Proceedings of the National Academy of Sciences*, *97*(17), 9591–9596. https://doi.org/10.1073/pnas.97.17.9591

Lier, E., & Kishk, A. (2005). A new class of dielectric-loaded hybrid-mode horn antennas with selective gain: Design and analysis by single mode model and method of moments. *IEEE Transactions on Antennas and Propagation*, *53*(1), 125–138. https://doi.org/10.1109/TAP.2004.840504

Lier, E. (2009). *Artificial dielectric antenna elements* (US7623085B1).

Lier, E. (2012). *Antenna array with metamaterial lens* (US8164531B2).

Lier, E. (2013). *Low index metamaterial* (US8466370B2).

Lier, E., & Katz, A. (2009). *Horn antenna, waveguide or apparatus including low index dielectric material* (US7629937B2).

Lier, E., & Katz, A. (2011). *Horn antenna, waveguide or apparatus including low index dielectric material* (US8072386B2).

Lu, M.-H., Weaver, M. S., Zhou, T. X., Rothman, M., Kwong, R. C., Hack, M., & Brown, J. J. (2002). High-efficiency top-emitting organic light-emitting devices. *Applied Physics Letters*, *81*(21), 3921–3923. https://doi.org/10.1063/1.1523150

Luk, V. N., Mo, G. C., & Wheeler, A. R. (2008). Pluronic Additives: A Solution to Sticky Problems in Digital Microfluidics. *Langmuir*, *24*(12), 6382–6389. https://doi.org/10.1021/la7039509

Peng, H., Zhu, X., Sun, J., Xie, Z., Xie, S., Wong, M., & Kwok, H.-S. (2005). Efficient organic light-emitting diode using semitransparent silver as anode. *Applied Physics Letters*, *87*(17), 173505. https://doi.org/10.1063/1.2115076

Summerton, J. (2016). History and Properties of Morpholino Antisense Oligos. *Journal of Drug Discovery, Development and Delivery*, *3*(1), 1019.

Summerton, J., & Weller, D. (1997). Morpholino antisense oligomers: Design, preparation, and properties. *Antisense and Nucleic Acid Drug Development*, *7*(3), 187–195. https://doi.org/10.1089/oli.1.1997.7.187

Wheeler, A. R., & Jebrail, M. J. (2011). *Digital microfluidic method for protein extraction by precipitation from heterogeneous mixtures* (US8053239B2).

Appendix B Data Construction

This section provides a comprehensive overview of the data sources, sample construction procedures, variable definitions utilized in the research, and descriptive statistics at the publication, patent, and firm-year levels. Table B1 provides a summary of variable definitions used in the analysis.

B.1 Data Sources

The primary data source used in this research is the Duke Innovation & SCientific Enterprises Research Network database (DISCERN, Arora et al. (2020)). DISCERN provides a match between patents and scientific publications to US-based publicly-traded firms between the years 1980 and 2015. For the construction of DISCERN, firm data were obtained from Compustat 2018 (firm-level data), ORBIS (subsidiaries), SDC Platinum (M&A activity), and WRDS CRSP (name changes). These data were matched to patent data from PatStat and scientific publication data from Web of Science (WoS). DISCERN is unique because it provides a match between patents and scientific publications to firms while considering firm structure and ownership changes. About a third of the firms in the sample changed their names within the sample years. Accounting for these changes improves the matching accuracy and provides a comprehensive baseline for studying firms' scientific and innovative activities.

Since creating DISCERN, several new data sources for scientific publications have become available. These sources provide key variables for the analyses presented in this work. Specifically, Dimensions was launched by Digital Science In January 2018.[56] It includes linked research information from over 128 million publications to over 99,000 journals, along with records of grants, datasets, patents, policy papers, clinical trials, and more (Herzog et al., 2020; Hook et al., 2018). Scientific publication data in Dimensions are sourced from Crossref and PubMed. Dimensions provides multiple data enhancements such as affiliation and researcher disambiguation, concept mapping, and more. Several recent works compared the coverage of Dimensions to previously available sources (such as WoS) and found it to have adequate coverage (e.g., Martín-Martín et al. (2021) and Singh et al. (2021)).

Along with Web of Science and Dimensions, I use Microsoft Academic Graph as a third source of scientific publications data. Microsoft launched the Academic Graph (MAG) in 2016 and provided open access to the complete dataset (Wang et al., 2020).[57] Importantly, MAG enables the construction of the citations count that is eventually linked to patent citations to science, obtained from Marx and Fuegi (2022).

Several complementary data sources are also included in the analysis. The American Men and Women of Science (AMWS) is a biographical directory of renowned North American scientists in the physical, biological, and related sciences. Entrants are scientists who have made significant contributions in their fields. Belenzon and Cioaca (2021) have acquired 17 electronic versions of the AMWS directory, covering editions published from 2005 through 2021. These editions include information about 240,800 living and deceased scientists. With the research assistance of Hansen Zhang, these data were matched to the DISCERN firms and Dimensions publications using textual similarity matching. I incorporate the matched data to study firms' hiring of scientists.

I incorporate data from the Google Patents database for analyses related to patents. Lastly, I add patent measures of private value from Kogan et al. (2017) and textual novelty from Kelly et al. (2021).

[56] https://www.dimensions.ai/

[57] In May 2021, Microsoft terminated the Academic Services project and future development of MAG. OurResearch, a non-profit organization founded in 2012, superseded MAG by releasing the OpenAlex project (Priem et al., 2022).

B.2 Sample Construction

The DISCERN data set includes 796,068 unique WoS records matched to 3,134 firms and published between 1980-2015.[58] First, I keep the 582,107 records of articles and proceedings papers and drop other types (such as books, editorial materials, letters, reviews, and meeting abstracts). Next, I rely on a fuzzy textual match (using TF-IDF) to link WoS records with their equivalent MAG identifiers. Using DOIs obtained from MAG, I join the sample with Dimensions records. The matched sample includes 463,027 records with both MAG and Dimensions identifiers.

I applied additional filters for the sample I use in the publication level analysis. In the earlier years of the sample, I encountered substantial issues of missing data fields and inconsistent quality of disambiguated author identifiers. To improve the quality and avoid truncation on both sides of the sample, I restrict the data to the years between 1990 and 2012. Next, I remove cases where journal volume and issue data are unavailable. I also remove outlier journals, such as journals with more than 24 issues per year (or less than 3) and issues with more than 100 articles (or less than 5). In addition, since the "first-in-first-out" allocation rule of accepted manuscripts into issues does not apply to conference proceedings and special journal issues, I limit the sample to standard journal publications. First, I drop conference proceedings, special issue publications, and supplementary materials based on indicators obtained from WoS. Second, to complement these indicators, I use the average H-index of all authors in a journal issue to locate some unmarked outlier issues and remove them. In the next step, since the econometric specification requires a log count of citations, I remove publications with zero citations. Lastly, I remove singleton cases due to the inclusion of fixed effects. My final sample at the publication level includes 164,495 observations (156,475 unique publications, since some publications are coauthored by researchers from multiple firms) matched to 1,527 firms.

Along with the main results at the publication level, I conduct several analyses at the patent and firm-year levels. To construct these data sets, I aggregate publication level measures and complement them with additional variables, such as patent quality measures (at the patent level) and financial outcomes (at the firm-year level).

B.3 Variable Construction

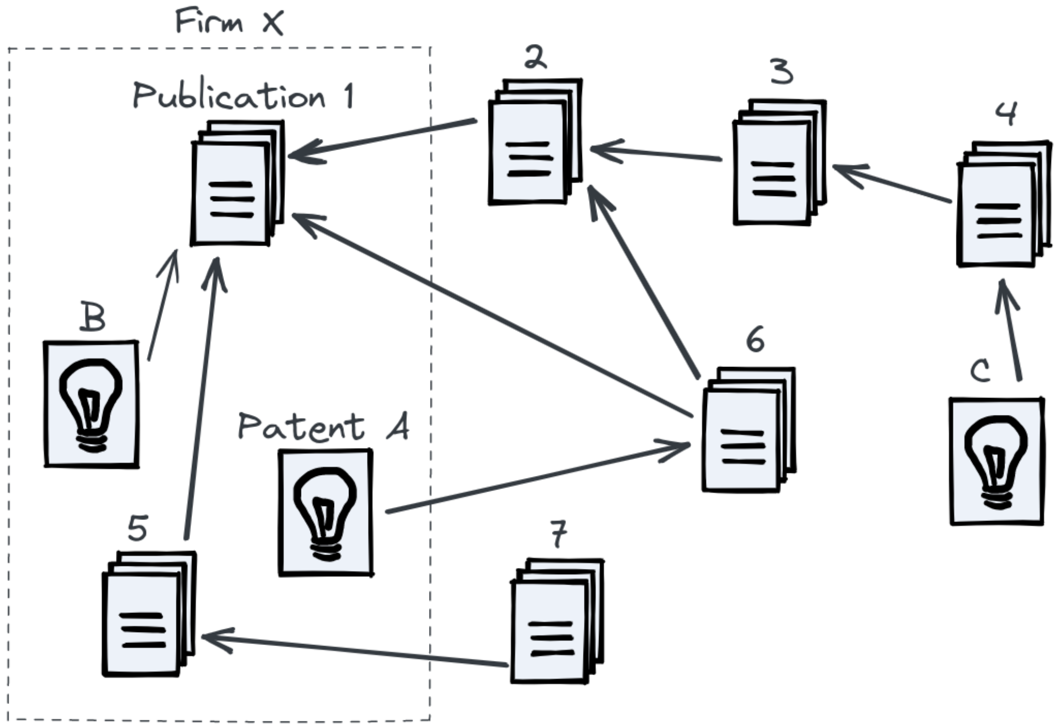
Below I detail the construction of the variables used in the analyses.

Follow-On Research I use citation data from MAG to count three generations of scientific citations to the focal publication. First, I restrict citing records to journal articles and conference proceedings published up to 2015. Next, I drop all citations within the firm by filtering the citations using my complete sample of corporate publications. I identify external direct citations and then rerun the match to identify second and third-generation citations. For each citing document, I keep the shortest citation route. I use the total count of citations to measure follow-on research originating outside the firm. Figure B1 illustrates the construction of the measure. Publication 1 is the focal paper published by firm X. Publications 2, 3, and 4 are counted as first, second, and third-generation follow-on research, respectively. I count publication 6 as a first-generation follow-on. Publication 5 is internal to the firm and, therefore, not counted as a follow-on. Publication 7 is not counted as a follow-on to publication 1 (however, it is counted as a direct follow-on to publication 5). Overall, publication 1 has four follow-on publications.

Patent citations to science (NPL citations) I use data from Marx and Fuegi (2022) to identify patents that cite scientific publications. I keep records related to USPTO patents with a confidence score equal to or greater than 3, filed up to 2015. I match NPL citations to my sample using the provided MAG identifiers. When considering patents citing follow-on research, I keep the shortest route between

[58] In some cases, publications are matched to multiple firms, so the total number of observations is 822,529.

Figure B1: Illustration of Follow-On Research



the patent and the focal publication. In Figure B1, patent A is an illustration of a patent by the same firm that cites follow-on research. Patent B directly uses science within the firm. Patent C is a patent by others that cite external research that is considered a follow-on to publication 1.

Author H-index I calculate temporal H-index measures both for the construction of the instrumental variable (other authors in the same journal issue) and as control variables (top author of focal publication). I use Dimensions journal, issue, and volume data to identify other publications in the same journal issue as the focal publication. To create the H-index measure, I use the Dimensions disambiguated researcher identifiers to first identify prior published works of all authors in the journal (published up to the focal publication year). Next, for each prior work, I identify all scientific citations received up to the same focal publication year. I count the citations for each prior work and then apply the H-index algorithm described in Section 3.2.3. I use the top H-index among the authors of the focal publication as a control variable. For the instrument, I sum the top two H-indexes among all authors in the journal issue after excluding the authors of the focal publication.

Future publications and patents by focal authors I use Dimensions disambiguated author identifiers to count subsequent scientific publications published by authors of the focal publication and related to the same firm. However, this method did not work for subsequent patents due to an incomplete match of inventors and scientific authors in the Dimensions data. Therefore, to identify patents by the focal authors, I conducted a textual match using the author's last name and first initial for inventors of USPTO patents assigned to the same firm.

AMWS Hiring I incorporate a match between AMWS, DISCERN and Dimensions publications. The data includes the employment years of AMWS scientists by the firms in my sample. Overall, 20,552 employed individuals are identified (26,385 records, as some individuals move between firms). However, my analysis requires identifying the employment of scientists whose work is related to the focal publication. To identify such relations, I restrict the sample to 6,673 individuals for whom I have scientific publication records. For these publications, I use the Dimensions concepts variable to identify granular research topics. The concepts are extracted by Dimensions from titles and abstracts of publications and their relevance is assigned using the pointwise mutual information algorithm against the publications field of research domain (FOR). I use a cutoff of 0.75 to identify highly relevant concepts.[59] Next, I match the scientist's employment data using the concepts from the focal publication and the concepts related to the scientist's works. If there is an overlap in concepts and the employment term begins after the focal publication year, I consider the individual's hiring by the firm as related to the focal publication.

Technological Leadership To identify firms' relative patenting capabilities, I first create a crosswalk between Dimensions scientific fields of research (FOR) and Cooperative Patent Classifications (CPC) at the 4-digit level. I use all USPTO patent NPL citation data from Marx and Fuegi (2022) and merge it with Dimensions using DOIs. Next, I aggregate the citations by FOR category and CPC. For each scientific category, I find the three most prevalent CPCs of citing patents. Using this crosswalk, I identify the set of related DISCERN patents filed in the same year as the focal publication. The ratio between a firm's patent count to the total patent count among all DISCERN firms is the measure of technological leadership related to the scientific domain of the focal publication.

University-industry collaborations I identify scientific publications and patents as university-industry collaborations (UIC) using Dimensions organization identifiers and the GRID data set. Records affiliated with the firm and with an organization of type "Education" are considered UIC. In an alternative specification, I identify UICs using the raw affiliation text and searching for 'univ|colleg|hosp'. The results of the analyses are similar for both specifications.

Patent-paper pairs I identify patent-paper pairs (PPP) using several steps. First, similarly to identifying future patents, I conduct a textual match of authors' last names and first initials, where both the focal publication and the patent are assigned to the same firm. Second, I restrict the patent filing date to be within two years of the focal publication date. Third, I require textual concept overlap using Dimensions concepts (as described above) between the patent and the focal publication. The resulting pairs are authored within the same firm by at least one shared author-inventor and share unique textual concepts.

Scientific concept prevalence First, I count Dimensions scientific concept appearances in non-corporate scientific publications by year. Second, I sum the concept counts in the previous three years for each concept related to the focal publication. Third, I aggregate the counts to the focal publication level. The resulting measure indicates if the concepts in the focal publication were prevalent in prior works by the scientific community.

Government funding I identify US government-funded publications using the funding acknowledgment field in Dimensions data. Similarly to the method described above for constructing the scientific concept prevalence, I count prior concept appearance only for government-funded publications. The ratio between government-funded concept appearances and the total concept appearance count is my measure of the availability of government funding for related scientific work.

[59]For example, a random list of concepts includes: target molecules, axial magnetic field, periodic solutions, amorphous silicon thin films, user attention, genetic interactions, polymer adsorption, coal-fired power plants, material removal rate, coverage algorithm, radical cation, laser pulse shape, nitrogen content, cell wall integrity, surface-enhanced Raman scattering, photophysical properties.

Table B1: Summary of Variables

Variable Name	Type	Measure of	Definition
Panel A: Publication level analysis			
Pr(Publication)	Indicator	Firm's subsequent scientific investments	Equals one if the corporate authors of focal publication publish a subsequent scientific paper, and zero otherwise.
Pr(Univ. Collab)	Indicator	Firm's subsequent direct ties with academics	Equals one if the corporate authors of focal publication publish a subsequent scientific paper with external academics, and zero otherwise.
Pr(Conference Proc.)	Indicator	Firm's subsequent participation in academics conferences	Equals one if the corporate authors of focal publication publish a subsequent conference proceeding, and zero otherwise.
Pr(AMWS Hire)	Indicator	Firm's hiring of a renowned scientist whose work is related to the focal publication	Equals one if the firm hires a related AMWS scientist in the years after the focal publication, and zero otherwise.
Pr(AMWS Award-winning Hire)	Indicator	Firm's hiring of an award-winning renowned scientist whose work is related to the focal publication	Equals one if the firm hires an award-winning related AMWS scientist in the years after the focal publication, and zero otherwise.
Pr(Patent, $\geq 3y$ gap)	Indicator	Firm's subsequent patenting outcomes	Equals one if any of the corporate authors of the focal publication file a subsequent patent at least three years after the focal publication year, and zero otherwise.
Pr(Patent, $\geq 10y$ gap)	Indicator	Firm's subsequent patenting outcomes	Equals one if any of the corporate authors of the focal publication file a subsequent patent at least ten years after the focal publication year, and zero otherwise.
Pr(Patent, Univ. Collab.)	Indicator	Firm's subsequent patenting with external academics	Equals one if any of the corporate authors of the focal publication file a subsequent patent that is co-assigned to a public research institution, and zero otherwise.
Pr(Patent Citation to Focal)	Indicator	Firm's subsequent patenting outcomes	Equals one if a patent assigned to the firm and filed after the focal publication year cites the focal publication, and zero otherwise.
Pr(Patent Citation to Focal or FO)	Indicator	Firm's subsequent patenting outcomes	Equals one if a patent assigned to the firm and filed after the focal publication year cites the focal publication or any of the follow-on publications, and zero otherwise.
Pr(Patent, $\geq 3y$ gap, Citing FO)	Indicator	Firm's subsequent patenting, follow-on research is an input	Equals one if any of the corporate authors of the focal publication file a subsequent patent that cites any of the follow-on research, and zero otherwise.
Pr(Patent, $\geq 3y$ gap, Not Citing FO)	Indicator	Firm's subsequent patenting, follow-on research provides quality validation	Equals one if any of the corporate authors of the focal publication file a subsequent patent that does not cite the follow-on research, and zero otherwise.
ln(Follow-On)	Continuous	External research that follows the focal publication	A logged count of external scientific publications that cite the focal publication, up to three generations away.
ln(Focal H-index)	Continuous	Prominence of the leading author of the focal publication	A logged H-index measure calculated at the focal publication year based on prior publications and citations.
ln(Top Two Researchers H-index)	Continuous	Prominence of the top two authors in the same journal issue of the focal publication	A logged sum of the H-index measures of the top two authors in a journal issue, after excluding the authors of the focal publication. Calculated at the focal publication year based on prior publications and citations.
Technological Leader	Indicator	Firm's relative patenting capability related to the focal publication	Above-median ratio between the number of related patents filed by the firm in the focal publication year over and related patents filed by all DISCERN firms.
Patent-Paper Pair (PPP)	Indicator	Possession of complementary IP rights	Equals one in the presence of a patent assigned to the same firm, invented by at least one of the authors of the focal publication, filed within two years of the focal publication year, and shares a textual concept.
University-Industry Collaboration (UIC)	Indicator	Knowledge Outsourcing	Equals one when at least one of the authors of the focal publication is affiliated with an educational institution, based on Dimensions GRID organization data.
High Scientific Concept Prevalence	Indicator	Nascent research domain	Equals one for above-median count of related scientific concept appearance in external publications in the three years prior to the focal publication year.
High Government Funding Availability	Indicator	Availability of government funding for the scientific community in related works	Equals one for above-median ratio between government-funded concept count and the total concept count in the three years prior to the focal publication year.
Panel B: Patent level analysis			
ln(NPL to Follow-On)	Continuous	External follow-on research related to the invention	Logged count of external scientific publications that cite the firm's scientific publications and are cited by the patent.

(continued on next page)

Table B1, continued

Variable Name	Type	Measure of	Definition
ln(NPL to Internal)	Continuous	Internal research related to the invention	Logged count of internal scientific publications that are cited by the patent.
ln(KPSS value)	Continuous	Private value of the patent	Estimate of private value of patent in real 2010 Dollars (Kogan et al., 2017).
ln(Word Count)	Continuous	Legal scope narrowness	Logged count of words in first claim of the patent.
Pr(KPST Breakthrough)	Indicator	Patent novelty	Top 10% of patents by textual novelty (Kelly et al., 2021).
Panel C: Firm-year level analysis			
Annual Publications	Count	Firm's scientific investments	Count of scientific publications by the firm, published in current year.
AMWS Scientist Employment	Count	Firm's scientific investments	Count of AMWS scientists employed by the firm.
Annual Patents	Count	Firm's patenting outcomes	Count of patents filed by the firm in current year.
ln(Follow-On Stock)	Continuous	Extent of external follow-on research	Logged stock of external scientific publications that cite the firm's publications (annual 15% depreciation rate applied).
ln(Firm's patent stock citing follow-on)	Continuous	Firm's use of follow-on research	Logged stock of firm's patents that cite follow-on research, aggregated by filing year (annual 15% depreciation rate applied).
ln(External patent stcok citing follow-on	Continuous	Other's use of follow-on research	Logged stock of external patents that cite follow-on research, aggregated by filing year (annual 15% depreciation rate applied).
ln(Tobin's Q)	Continuous	Firm's financial performance	Logged market value over assets.
ln(Future Follow-On)	Continuous	Future follow-on research	Logged stock of external research that cites firm's publications and is published after current year (one, two or three generations of citations).
ln(Future Internal NPL to follow-on)	Continuous	Future use of follow-on research by the firm	Logged count of firm's patent citations to follow-on research, by patents filed after current year.
ln(Future external NPL to follow-on)	Continuous	Future use of follow-on research by others outside the firm	Logged count of external patent citations to follow-on research, by patents filed after current year.
Publications/R&D	Continuous	Firm's scientific publication stock, scaled by R&D investments	The ratio of publication stock and R&D investment stock.
Patents/R&D	Continuous	Firm's patenting stocks, scaled by R&D investments	The ratio of patent stock and R&D investment stock.
R&D/Assets	Continuous	R&D intensity	Firm's R&D stock over asset stock.

B.4 Supplementary Descriptive Statistics

Table B2 provides descriptive statistics for the patent sample. Table B3 provides descriptive statistics for the patent sample.

Table B2: Descriptive Statistics for Patent Sample

Variable	Missing	Mean	SD	Min	p25	p50	p75	Max
Grant Year	0	2,004.1	9.1	1980	1998	2006	2012	2015
NPL to Follow-On	0	10.5	27.6	0	1	2	7	954
Upstream Internal pubs	0	16.1	73.2	0	0	0	2	2205
First Claim Word count	0	171.6	192.0	10	92	139	203	18844
KPSS Patent Value	59,643	18.1	43.7	0	3	7	16	3522
KPST Novelty	154,595	0.2	0.4	0	0	0	0	1

This table provides summary statistics for the main variables used in the econometric analysis at the patent level. The sample is based on the DISCERN database and includes patents with at least one reference to a scientific publication (NPL) and assigned to U.S. publicly-traded firms between the years 1980-2015. Note that from the total of 492,871 DISCERN patents with NPL citations, 8,537 observations are dropped due to multiple firm assignment or missing variables.

Table B3: Descriptive Statistics for the Firm-Year Panel

Variable	Mean	SD	Min	p25	p50	p75	Max
Year	1,999.2	9.2	1981	1992	2000	2007	2015
Annual Publications	9.0	57.1	0	0	0	2	1,590
Employed AMWS Scientists	5.3	34.3	0	0	0	1	1,014
Employed AMWS Scientists, Award-winning	1.9	9.5	0	0	0	1	170
Annual Patents	27.0	146.8	0	0	1	8	8,842
Follow-On Research Stock	36,797.2	234,416.8	0	0	4	1,191	6,784,316
Firm's patent stock citing FO	14.7	185.3	0	0	0	0	15,456
External patent stock citing FO	1,775.1	10,630.9	0	0	0	22	325,990
Tobin's Q	24.3	448.8	0	1	2	4	63,621
Future FO Research (1st gen.)	1,574.0	11,372.3	0	0	3	141	292,574
Future FO Research (1,2 gen.)	32,492.6	223,899.5	0	0	25	2,113	4,853,391
Future FO Research (1-3 gen.)	55,632.7	333,843.4	0	0	0	1,122	6,526,805
Publication Stock	86.5	526.8	0	0	2	17	12,251
Patent Stock	141.5	777.7	0	2	8	42	31,551
R&D Stock	463.2	2,573.6	0	6	29	134	58,450
Asset Stock	1,949.2	10,946.5	0	12	77	572	355,367

*Notes:*This table provides summary statistics for the variables used in the econometric analysis at the firm-year level. The data is based on the DISCERN database of publications by U.S.-based publicly-owned firms between 1980 and 2015. The sample includes 43,303 firm-year observations.

References

Arora, A., Belenzon, S., & Sheer, L. (2020). DISCERN: Duke Innovation & SCientific Enterprises Research Network. https://doi.org/10.5281/zenodo.4320782

Belenzon, S., & Cioaca, L. C. (2021). *Guaranteed Public Demand and Corporate Scientific Research* (Working Paper No. 28644). National Bureau of Economic Research. https://doi.org/10.3386/w28644

Herzog, C., Hook, D., & Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. *Quantitative Science Studies*, *1*(1), 387–395. https://doi.org/10.1162/qss_a_00020

Hook, D. W., Porter, S. J., & Herzog, C. (2018). Dimensions: Building Context for Search and Evaluation. *Frontiers in Research Metrics and Analytics*, *3*. https://doi.org/10.3389/frma.2018.00023

Kelly, B., Papanikolaou, D., Seru, A., & Taddy, M. (2021). Measuring Technological Innovation over the Long Run. *American Economic Review: Insights*, *3*(3), 303–320. https://doi.org/10.1257/aeri.20190499

Kogan, L., Papanikolaou, D., Seru, A., & Stoffman, N. (2017). Technological Innovation, Resource Allocation, and Growth. *The Quarterly Journal of Economics*, *132*(2), 665–712. https://doi.org/10.1093/qje/qjw040

Martín-Martín, A., Thelwall, M., Orduna-Malea, E., & Delgado López-Cózar, E. (2021). Google Scholar, Microsoft Academic, Scopus, Dimensions, Web of Science, and OpenCitations' COCI: A multidisciplinary comparison of coverage via citations. *Scientometrics*, *126*(1), 871–906. https://doi.org/10.1007/s11192-020-03690-4

Marx, M., & Fuegi, A. (2022). Reliance on Science in Patenting. https://doi.org/10.5281/zenodo.6468792

Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. https://doi.org/10.48550/arXiv.2205.01833

Singh, V. K., Singh, P., Karmakar, M., Leta, J., & Mayr, P. (2021). The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. *Scientometrics*. https://doi.org/10.1007/s11192-021-03948-5

Wang, K., Shen, Z., Huang, C., Wu, C.-H., Dong, Y., & Kanakia, A. (2020). Microsoft Academic Graph: When experts are not enough. *Quantitative Science Studies*, *1*(1), 396–413. https://doi.org/10.1162/qss_a_00021

Appendix C Additional Results

C.1 Supplementary Descriptive Analysis

C.1.1 Patents using internal and follow-on research

Following the analysis in Section 4, a similar result emerges by classifying corporate patents based on their use of science. Table C1 presents descriptive statistics at the NPL citation and patent levels. According to columns 1 and 2, out of 6.6 million citations from corporate patents to science, only 2% are to publications by the same firm. An additional 18% are to follow-on research, and the remaining citations are to external research unrelated to the firm. Columns 3-6 aggregate NPL citations to the patent level. Out of 492,871 science-based patents,[60] only 12% directly cite a scientific publication by the firm. An additional 22.5% do not cite internal science but cite external follow-on research. After accounting for truncation, I find that more than 40% of corporate science-based patents are directly or indirectly related to firms' contributions to public research. Lastly, among patents that directly cite internal publications (columns 7-8), about 35% also cite external follow-on research.

C.1.2 Time Trends in the Use of Follow-On Research

Figure C1 complements Figure 1 and Table 1. It presents time trends in the distribution of corporate publications by publication year and eventual direct and indirect patent citations. Note that the decline in later years is due to truncation of the data in 2015. The figure suggests that, at least up to the early 2000s, there is no clear change in the rate of direct and indirect patent citations to firms' publications throughout the sample years. A breakdown of these trends by field (available upon request) provides similar results, with slight variation across fields.

C.2 Supplementary Regression Analysis

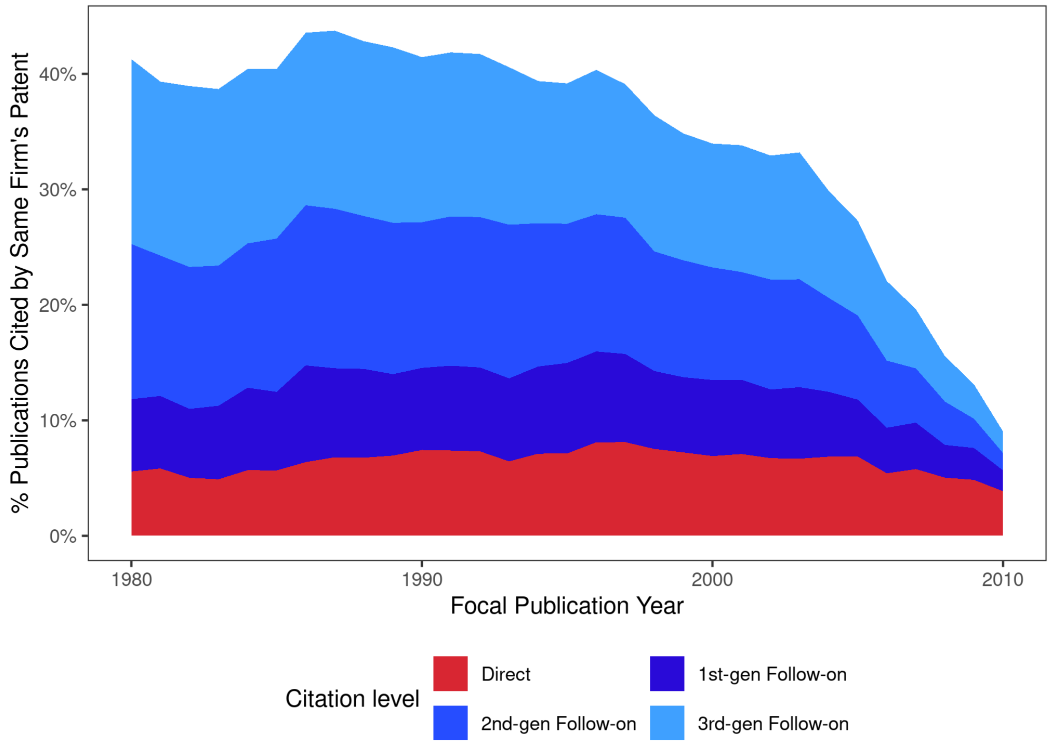
C.2.1 Poisson Estimation of Count Models

Typically, models with counts as the dependent variable are best estimated using Poisson regressions (Wooldridge, 2010). Methods are available for the estimation of two-stage Poisson (e.g., ivpoisson in Stata), and the estimation of Poisson with fixed effects (e.g., ppmlhdfe in Stata, fepois from the fixest package in R). However, there is currently no accepted implementation of a two-staged Poisson regression that allows the inclusion a large number of fixed effects. Therefore, for my main analyses I present a binary outcome (equal to one for a positive count) and a linear probability model estimated with standard OLS and 2SLS. Tables C2 and C3 present supplementary analyses for the baseline results using Poisson fixed effect estimations of count models. While these regressions do not account for endogeneity, the estimated positive correlations provide additional support for the results presented in the paper.

In Poisson estimation, observations are automatically dropped to avoid singletons and separation. One source of separation is a constant dependent variable within a fixed effect group. Table C4 and C5 present OLS and 2SLS results (specifications corresponding to the baseline results) for the observations that remained in the Poisson samples above. Overall, these results are similar to the results presented in the paper.

[60] I define a science-based patent as a patent with at least one citation to a scientific publication.

Figure C1: Time Trends in the Use of Follow-On Research



Note: This figure shows time trends in the distribution of corporate publications by publication year and eventual direct and indirect patent citations. Publications that are directly cited are publications for which there is patent by the same firm that cites them. Publications that are indirectly cited are publications for which there is external follow-on research that is eventually cited by a patent by the same firm. A patent is counted once for each focal publication based on the shortest citation route. The decline in later years is due to truncation of the data in 2015.

C.2.2 Heterogeneity in Subsequent NPL Citations

Table 8 explores heterogeneity in the estimations of the effect of follow-on research on the likelihood of subsequent patent by the focal authors. In the baseline results (Table 4), I also present an analysis of the effect of follow-on research on the likelihood that subsequent patents by the firm (not restricted to the focal authors) will cite the focal publication. Table C6 presents a corresponding heterogeneity analysis. Overall, the results are qualitatively similar but weaker in statistical significance.

C.2.3 Firms' Conference Proceedings Sample

Firms' conference proceedings are dropped from the baseline sample because they do not follow the same assignment procedures into journal issues (see Section B.2 for details). However, proceedings are an important part of firms' scientific output, especially in some specific fields (e.g., computer science). Table C7 presents an analysis of firms' conference proceedings that were dropped from the main sample. The positive estimated correlations are in line with the baseline results in the paper.

C.2.4 Variation by Field and Industry

Table C8 presents estimation results for subsamples based on firms' main industries. Industry classifications are based on 2-digit SIC codes. Note that the subsamples not equal in size. Since the subsamples are much smaller than the full sample, in none of them the instrument has enough power to predict follow-on research. I therefore present OLS regressions. The results indicate that in most industries, there is a positive and statistically significant relation between follow-on research and subsequent scientific publication and patenting by the focal authors. Focusing on the chemical industry (including medicine), I find a significantly stronger relation compared to other industries (columns 6 and 13).

Table C9 presents estimation results for subsamples based on publications' assigned research field. Fields are determined by Dimensions.ai based on Australian and New Zealand Standard Research Classification (ANZSRC) 2020. Similar to the case of industries, subsamples are too small for 2SLS estimation. The results indicate a positive relation to subsequent publishing and patenting across different fields. However, differences across fields are not statistically significant (e.g., columns 6 and 13).

C.2.5 Corporate vs Academic Follow-On Research

An interesting extension to the baseline results would be to explore how different sources of follow-on affect the focal firm. For example, it seems likely that follow-on research that originates for universities will have a different effect than follow-on research that originates from other firms. I attempt to explore these differences in Table C10. I split the count of follow-on research by source, based on Dimensions.ai classification of institution type. Academic follow-on research is research that originates from academic and other non-private institutions. Corporate follow-on research originates from other firms.

In both OLS and 2SLS regressions, I find very similar coefficients (with the exception of columns 3 and 5). Note, however, that in this case it is hard to defend the exclusion restriction of the instrument. For example, in column 6, while the instrument increases citations to corporate follow-on research, it also affects non-corporate follow-on research. There is not enough variation in the data to separately include both variables in the regression. It therefore seems that a different design is required in order to explore this source of heterogeneity.

Table C1: Firms' Patent Citations to Own Scientific Publications

	NPL Citations		Patents					
			All		Granted 2000-2015		Directly Citing	
	Count (1)	Percent (2)	Count (3)	Percent (4)	Count (5)	Percent (6)	Count (7)	Percent (8)
Direct citation to a firm's publication	132,568	2.02%	58,310	11.83%	46,038	13.04%	58,310	11.83%
Indirect citation to a firm's publication (citing follow-on research)								
1st Generation	251,331	3.82%	38,155	7.74%	33,542	9.50%	16,022	27.50%
2nd Generation	509,621	7.75%	41,645	8.45%	37,620	10.66%	13,021	22.33%
3rd Generation	445,470	6.78%	31,169	6.32%	27,745	7.86%	11,739	20.01%
Any	1,206,422	18.36%	110,969	22.51%	98,907	28.02%	20,375	34.95%
Citing a firm's publication (directly or indirectly)	1,338,990	20.37%	169,279	34.35%	144,945	41.6%	58,310	100%
Not Citing a firm's publication	5,233,655	79.63%	323,592	65.65%	208,059	58.94%	0	0%
Total	6,572,645	100.00%	492,871	100.00%	353,004	100.00%	58,310	100%

This table presents summary statistics of citations from firms' patents to scientific publications (NPL). The dataset includes 492,871 patents with a total of 6.6 million NPL citations. In columns 1-6, a patent is categorized by the shortest route to an internal scientific publication. 58,310 (12%) patents directly cite a scientific publication published by the same firm. An additional 110,969 (22.5%) patents cite external scientific publications that have an internal publication as an upstream reference, up to the 3rd generation of references. 323,592 patents (65.5%) do not have a citation to an upstream internal scientific publication. To account for truncation, columns 5 and 6 present a subset of patents published after the year 2000. Among these patents, about 41% cite a firm's publication either directly or indirectly. Columns 7-8 only include patents that directly cite internal scientific publications. Among these patents, all citation routes are considered and patents are classified based on whether they include a citation to different levels of follow-on research. The results indicate that about 35% of directly-citing patents also cite external follow-on research.

Table C2: Follow-On Research and Firms' Investments in Science

	Subsequent Scientific Publications by Focal Authors				Hiring of Renowned Scientists (AMWS)	
	Publication Count	Pub. Count, Univ. Collab.		Pub. Count, Conf. Proc.	Hire Count	Award-winning Hire Count
	PPML	PPML	PPML	PPML	PPML	PPML
	(1)	(2)	(3)	(4)	(5)	(6))
ln(Follow-On)	0.053***	0.067***	0.011***	0.048***	0.072***	0.066***
	(0.006)	(0.007)	(0.002)	(0.016)	(0.004)	(0.005)
ln(Future Pubs)			1.088***			
			(0.014)			
ln(Focal H-Index)	0.221***	0.320***	0.116***	0.184***	0.029***	0.023
	(0.019)	(0.021)	(0.011)	(0.033)	(0.009)	(0.019)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Journal-Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	142,610	137,294	137,294	66,036	61,573	15,622
Avg. DV	20.943	8.690	8.690	2.390	0.780	0.215
Pseudo R^2	0.489	0.489	0.869	0.549	0.553	0.333

Notes: This table accompanies table 3 and reports estimation results of corresponding count models. Estimation is conducted using Poisson pseudo maximum likelihood (Correia et al., 2020). Observations are automatically dropped to avoid singletons and separation. Clustered (Firm) standard-errors in parentheses. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

Table C3: Follow-On Research and Firms' Patenting Outcomes

	Subsequent Patents by Focal Authors			Subsequent Firms' Patents	
	Patent Count $\geq$3y gap	Patent Count $\geq$10y gap	Patent Count Univ. Collab.	Citation Count to Focal	Citation Count to Focal or FO
	PPML	PPML	PPML	PPML	PPML
	(1)	(2)	(3)	(4)	(5)
ln(Follow-On)	0.028**	0.021**	0.075***	0.381***	0.718***
	(0.012)	(0.010)	(0.021)	(0.050)	(0.024)
ln(Focal H-Index)	0.051**	0.032	0.425***	-0.071	-0.031
	(0.026)	(0.025)	(0.081)	(0.066)	(0.036)
Firm FE	Yes	Yes	Yes	Yes	Yes
Journal-Year FE	Yes	Yes	Yes	Yes	Yes
Observations	133,053	63,156	41,138	65,229	119,144
Avg. DV	7.952	4.681	0.463	0.699	17.506
Pseudo R^2	0.584	0.567	0.588	0.523	0.855

Notes: This table accompanies table 4 and reports estimation results of corresponding count models. Estimation is conducted using Poisson pseudo maximum likelihood Correia et al. (2020). Observations are automatically dropped to avoid singletons and separation.
Clustered (Firm) standard-errors in parentheses. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

Table C4: The Effect of Follow-On Research on Firms' Investments in Science, Filtered Sample

	Subsequent Scientific Publications by Focal Authors						Hiring of Renown Scientists (AMWS)			
	Pr(Publication)		Pr(Univ. Collab.)		Pr(Conference Proc.)		Pr(Hire)		Pr(Award-winning Hire)	
	OLS	2SLS	OLS	2SLS	OLS	2SLS	OLS	2SLS	OLS	2SLS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
ln(Follow-On)	0.009***	0.134***	0.013***	0.156***	0.004**	0.076	0.009***	0.073*	0.009***	0.200**
	(0.002)	(0.050)	(0.002)	(0.056)	(0.002)	(0.055)	(0.002)	(0.038)	(0.003)	(0.090)
ln(Focal H-Index)	0.002	-0.033**	0.038***	-0.002	0.012***	-0.010	0.005**	-0.013	0.001	-0.055**
	(0.004)	(0.015)	(0.004)	(0.017)	(0.005)	(0.019)	(0.002)	(0.011)	(0.004)	(0.026)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Journal-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	142,610	142,610	137,294	137,294	66,036	66,036	61,573	61,573	15,622	15,622
Avg. DV	0.584	0.584	0.521	0.521	0.232	0.232	0.208	0.208	0.182	0.182
First Stage F-stat		30.991		27.932		13.187		22.100		5.220
Adjusted R^2	0.258	-0.364	0.239	-0.407	0.220	-0.201	0.227	-0.210	0.282	-1.341

Notes: This table is a filtered version of Table 3. The samples include observations that are kept after PPML estimation (as presented in table C2). Observations are automatically dropped to avoid singletons and separation.
Clustered (Firm) standard-errors in parentheses. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

Table C5: The Effect of Follow-On Research on Firms' Patenting Outcomes, Filtered Sample

	Subsequent Patents by Focal Authors						Subsequent Firms' Patents			
	Pr(Patent, $\geq$3y gap)		Pr(Patent, $\geq$10y gap)		Pr(Patent, Univ. Collab.)		Pr(Citation to Focal)		Pr(Citation to Focal or FO)	
	OLS (1)	2SLS (2)	OLS (3)	2SLS (4)	OLS (5)	2SLS (6)	OLS (7)	2SLS (8)	OLS (9)	2SLS (10)
ln(Follow-On)	0.008***	0.061	0.005***	0.166**	0.008***	0.092	0.026***	0.110**	0.096***	0.118***
	(0.001)	(0.050)	(0.001)	(0.081)	(0.001)	(0.072)	(0.002)	(0.055)	(0.005)	(0.037)
ln(Focal H-Index)	0.013***	-0.002	0.005	-0.033*	0.029***	0.008	-0.019***	-0.042***	0.002	-0.004
	(0.002)	(0.014)	(0.003)	(0.019)	(0.006)	(0.017)	(0.003)	(0.016)	(0.002)	(0.011)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Journal-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	133,053	133,053	63,156	63,156	41,135	41,135	65,229	65,229	119,144	119,144
Avg. DV	0.490	0.490	0.367	0.367	0.130	0.130	0.141	0.141	0.417	0.417
First Stage F-stat		27.422		12.398		7.118		16.155		42.389
Adjusted R^2	0.274	-0.182	0.246	-0.603	0.202	-0.286	0.105	-0.235	0.432	0.029

Notes: This table is a filtered version of Table 4. The samples include observations that are kept after PPML estimation (as presented in table C3). Observations are automatically dropped to avoid singletons and separation.
Clustered (Firm) standard-errors in parentheses. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

Table C6: Heterogeneity in Subsequent Patenting (NPL Citations)

	Pr(Subsequent Firm Patent Citing Focal Publication)							
	OLS (1)	2SLS (2)	OLS (3)	2SLS (4)	OLS (5)	2SLS (6)	OLS (7)	2SLS (8)
Complementary IP Rights								
ln(Follow-On) × PPP	0.020*** (0.001)	0.023*** (0.005)						
PPP	0.066*** (0.004)	0.052*** (0.010)						
Knowledge Outsourcing								
ln(Follow-On) × UIC			-0.007*** (0.001)	-0.005 (0.004)				
UIC			-0.042*** (0.004)	-0.043*** (0.004)				
Scientific Concept Prevalence								
ln(Follow-On) × Low Prev.					0.000 (0.001)	0.008* (0.004)		
Low Prevalence					0.004*** (0.001)	0.001 (0.003)		
Government Funding Availability								
ln(Follow-On) × Govt. Funding							0.001 (0.001)	0.008* (0.005)
Govt. Funding							0.004** (0.002)	-0.002 (0.005)
ln(Follow-On)	0.005*** (0.001)	0.045** (0.023)	0.015*** (0.001)	0.044* (0.025)	0.011*** (0.001)	0.039 (0.025)	0.011*** (0.001)	0.040 (0.025)
ln(Focal H-Index)	-0.006*** (0.001)	-0.017*** (0.006)	-0.001 (0.001)	-0.009 (0.008)	-0.008*** (0.001)	-0.017** (0.008)	-0.008*** (0.001)	-0.017** (0.008)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Journal-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	164,495	164,495	164,495	164,495	164,495	164,495	164,495	164,495
Avg. DV	0.036	0.036	0.056	0.056	0.056	0.056	0.056	0.056
First Stage F-stat		15.215		14.452		14.423		14.878
Adjusted R^2	0.077	-0.287	0.078	-0.208	0.071	-0.228	0.071	-0.224

Notes: This table accompanies table 8. In this table, the dependent variable is an indicator for a subsequent patent by the focal firm that cites the focal publication.
Clustered (Firm) standard-errors in parentheses. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

Table C7: Corporate Conference Proceedings

	Subsequent Publications	Pr(Subsequent Publication)	Subsequent Patents	Pr(Subsequent Patent)
	PPML	OLS	PPML	OLS
	(1)	(2)	(3)	(4)
ln(Follow-On)	0.056***	0.003	0.038***	0.005***
	(0.011)	(0.003)	(0.006)	(0.002)
ln(Focal H-Index)	0.299***	-0.004	0.084***	0.005
	(0.043)	(0.010)	(0.020)	(0.005)
Firm FE	Yes	Yes	Yes	Yes
Conference FE	Yes	Yes	Yes	Yes
Observations	21,276	21,276	23,580	23,580
Avg. DV	12.903	0.442	33.922	0.790
Adjusted R^2		0.267		0.245
Pseudo R^2	0.609		0.617	

Notes: This table presents the baseline results for a sample of corporate conference proceedings that are dropped from the main analysis. The data consists of a pooled cross-section of proceedings by U.S.-based publicly-owned firms, published between 1990 and 2012 (Arora et al., 2020). Follow-on research is the total count of three generations of citations to the focal publication from outside the firm. The dependent variables are counts and indicators for subsequent scientific publications (columns 1-2) and patents (columns 3-4) by the corporate authors of the focal papers. All regressions include a control for the highest H-index among the authors of the focal publication, as well as firm and conference fixed effects.

Clustered (Firm) standard-errors in parentheses. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

Table C8: Variation by Main Industry

	Pr(Subsequent Scientific Publication by Focal Authors)							Pr(Subsequent Patent by Focal Authors, $\geq$ 3y gap)						
	Chem OLS (1)	Elec OLS (2)	Instr OLS (3)	Serv OLS (4)	Other OLS (5)	All OLS (6)	All 2SLS (7)	Chem OLS (8)	Elec OLS (9)	Instr OLS (10)	Serv OLS (11)	Other OLS (12)	All OLS (13)	All 2SLS (14)
ln(Follow-On)	0.011***	0.003	0.001	0.010***	0.004**	0.004**	0.106**	0.008***	0.007***	0.001	0.012***	0.007***	0.006***	0.087*
	(0.002)	(0.003)	(0.005)	(0.003)	(0.002)	(0.002)	(0.049)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.051)
ln(Follow-On) × Chemicals						0.008***	0.052						0.004*	-0.048
						(0.002)	(0.038)						(0.002)	(0.041)
ln(Focal H-Index)	-0.009***	0.006	0.018	0.009	0.026***	0.002	-0.033**	0.012***	0.005	0.034***	-0.000	0.015*	0.012***	-0.004
	(0.004)	(0.007)	(0.018)	(0.007)	(0.007)	(0.003)	(0.014)	(0.002)	(0.006)	(0.006)	(0.005)	(0.009)	(0.002)	(0.013)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Journal-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	88,174	15,875	13,470	14,692	15,248	164,495	164,495	88,174	15,875	13,470	14,692	15,248	164,495	164,495
Avg. DV	0.623	0.421	0.300	0.408	0.371	0.507	0.507	0.352	0.549	0.583	0.358	0.427	0.397	0.397
First Stage F-stat							15.062							15.062
Adjusted R^2	0.312	0.299	0.380	0.348	0.338	0.345	-0.420	0.327	0.342	0.475	0.318	0.404	0.348	-0.247

Notes: This table presents variation in the baseline estimation results by firms' main industry classifications. The analysis corresponds to Tables 3 and 4. Industry classification is based on 2-digit SIC: Columns 1 and 8 include "Chemicals And Allied Products" (SIC 28); Columns 2 and 9 include "Industrial And Commercial Machinery And Computer Equipment" (SIC 35) and "Electronic And Other Electrical Equipment And Components, Except Computer Equipment" (SIC 36); Columns 3 and 10 include "Transportation Equipment" (SIC 37) and "Measuring, Analyzing, And Controlling Instruments; Photographic, Medical And Optical Goods; Watches And Clocks" (SIC 38); Column 4 and 11 include "Business Services" (SIC 73) and "Engineering, Accounting, Research, Management, And Related Services" (SIC 87); Columns 5 and 12 include all other firms. Columns 6, 7, 13 and 14 include the complete sample and Chemicals is an indicator for SIC 28.

Clustered (Firm) standard-errors in parentheses. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

Table C9: Variation by Research Field

	Pr(Subsequent Scientific Publication by Focal Authors)							Pr(Subsequent Patent by Focal Authors, $\geq$ 3y gap)						
	Med OLS (1)	Chem OLS (2)	Eng & ICT OLS (3)	Math & Phys OLS (4)	Other OLS (5)	All OLS (6)	All 2SLS (7)	Med OLS (8)	Chem OLS (9)	Eng & ICT OLS (10)	Math & Phys OLS (11)	Other OLS (12)	All OLS (13)	All 2SLS (14)
ln(Follow-On)	0.014***	0.002	0.005***	0.002	0.012***	0.006***	0.103**	0.009***	0.006***	0.006***	0.004	0.008**	0.007***	-0.004
	(0.002)	(0.002)	(0.002)	(0.003)	(0.004)	(0.002)	(0.044)	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.001)	(0.059)
ln(Follow-On) $\times$ Chem & Med						0.003	0.048						0.001	0.196
						(0.002)	(0.115)						(0.002)	(0.154)
ln(Focal H-Index)	-0.011***	-0.009	0.016***	0.013*	0.018**	0.002	-0.033*	0.010***	0.013**	0.013***	0.026***	0.018**	0.012***	-0.017
	(0.003)	(0.006)	(0.006)	(0.007)	(0.009)	(0.003)	(0.018)	(0.003)	(0.005)	(0.004)	(0.008)	(0.008)	(0.002)	(0.018)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Journal-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	78,041	25,161	47,958	11,287	7,813	164,495	164,495	78,041	25,161	47,958	11,287	7,813	164,495	164,495
Avg. DV	0.594	0.584	0.372	0.358	0.370	0.507	0.507	0.310	0.512	0.508	0.471	0.275	0.397	0.397
First Stage F-stat							3.379							3.379
Adjusted R^2	0.308	0.409	0.291	0.404	0.341	0.344	-0.412	0.301	0.367	0.350	0.334	0.299	0.348	-0.524

Notes: This table presents variation in the baseline estimation results by publications' research field. The analysis corresponds to Tables 3 and 4. Research fields are determined by Dimensions.ai based on Australian and New Zealand Standard Research Classification (ANZSRC) 2020. Columns 1 and 8 include "Biological Sciences" (FOR 31), "Biomedical and Clinical Sciences" (FOR 32) and "Health Sciences" (FOR 42); Columns 2 and 9 include "Chemical Sciences" (FOR 34); Columns 3 and 10 include "Engineering" (FOR 40) and "Information and Computing Sciences" (FOR 46); Columns 4 and 11 include "Mathematical Sciences" (FOR 49) and "Physical Sciences" (FOR 51); Columns 5 and 12 include all other publications. Columns 6, 7, 13 and 14 include the complete sample and "Chem & Med" is an indicator for FOR codes 32, 34, 31 and 34.
Clustered (Firm) standard-errors in parentheses. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

Table C10: Corporate vs Academic Follow-On Research

	Pr(Subsequent Scientific Publication by Focal Authors)						Pr(Subsequent Patent by Focal Authors, $\geq$ 10y gap)					
	OLS	2SLS	OLS	2SLS	OLS	2SLS	OLS	2SLS	OLS	2SLS	OLS	2SLS
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
ln(All Follow-On)	0.008***	0.121**					0.003***	0.095**				
	(0.001)	(0.047)					(0.001)	(0.040)				
ln(Academic Follow-On)			0.007***	0.126***					0.003***	0.100**		
			(0.001)	(0.048)					(0.001)	(0.041)		
ln(Corporate Follow-On)					0.014***	0.126**					0.003***	0.113**
					(0.002)	(0.057)					(0.001)	(0.048)
ln(Focal H-Index)	0.002	-0.030**	0.002	-0.032**	-0.001	-0.024**	0.002	-0.024**	0.002	-0.025**	0.003	-0.020**
	(0.003)	(0.014)	(0.003)	(0.014)	(0.003)	(0.012)	(0.002)	(0.011)	(0.002)	(0.011)	(0.002)	(0.010)
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Journal-Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	164,495	164,495	164,127	164,127	154,583	154,583	164,495	164,495	164,127	164,127	154,583	154,583
Avg. DV	0.507	0.507	0.507	0.507	0.515	0.515	0.143	0.143	0.143	0.143	0.147	0.147
First Stage F-stat		29.528		27.987		33.607		29.528		27.987		33.607
Adjusted R^2	0.344	-0.377	0.344	-0.398	0.347	-0.317	0.351	-0.486	0.351	-0.518	0.352	-0.508

Notes: This table presents an analysis of follow-on research split by the type of source organization. The analysis corresponds to the baseline results presented in Tables 3 and 4. Columns 1, 2, 7 and 8 replicate the baseline results. In columns 3, 4, 9 and 10, the independent variable of interest is a count of three generations of follow-on research that originates from academic and other non-private institutions. In columns 5, 6, 11 and 12, the independent variable of interest is a count of three generations of follow-on research that originates from other firms. Institution type is determined based on a classification of authorship affiliations provided by the Dimensions dataset. Clustered (Firm) standard-errors in parentheses. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

Appendix D Discussion of Instrumental Variable

In this section, I discuss the main assumptions underlying my instrumental variable approach for identifying the effects of external follow-on research on firms' innovation outcomes.

D.1 Instrument Relevance

The instrument's relevance relies on the notion of peer effects across publications grouped together in physical journal issues. The elements of this process are discussed in detail in section 5.1. In short, until academic readership moved online in the early 2000s, academics accessed most scientific publications by walking to their institution's library and checking out physical journal issues. As a result, publications that were grouped in the same journal issue with a publication by a prominent researcher were likely to be circulated more often than others. Therefore, the level of attention to journal publications could have varied irrespectively of the content and quality of a given paper. I argue that serendipitous increases in academic attention sometimes translate into meaningful follow-on research. This research can then be observed through the number of citations the focal publication received.

Table D1 reports the first-stage coefficient estimates for the instrument's relevance. First, I use the top one H-index among all other authors in the same journal issue. Next, I consider the sum of the H-indexes of the top two authors. Since the predictive power is stronger under this specification, I chose it as the instrument across all analyses in the paper. Figure D1 presents a corresponding binned scatterplot. In addition to the chosen specification, I report in Table D1 coefficient estimates of alternative specifications for the instrument. First, I use the count of the authors' publications up to the year before the focal publication year. Second, I use the citation-weighted measure of the same publications. In all cases, I find strong evidence for the relevance of the instrument for the count of follow-on citations.

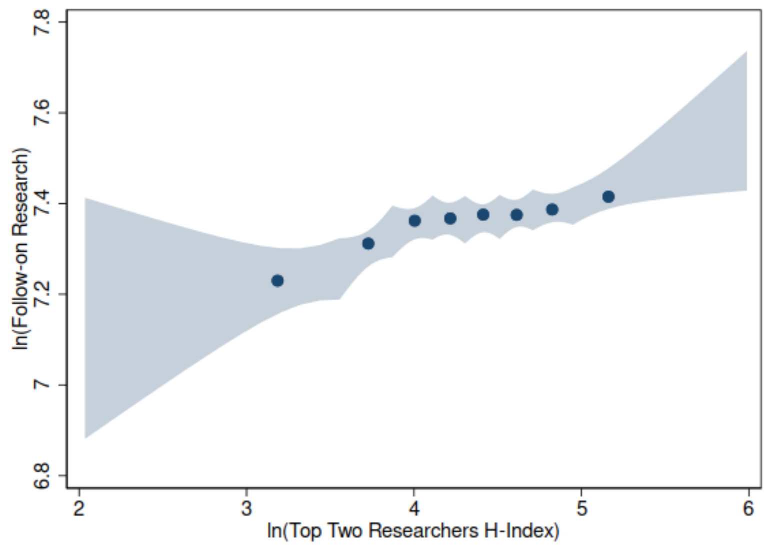
To further support the mechanism that drives the instrument relevance, I explore the time trend of coefficient estimates. Figure D2 presents the coefficient estimates of the interaction between the instrumental variable and 2-year indicators. As expected, the correlation between the instrument and follow-on citations was stronger during the 1990s, before academic readership moved online. Starting in the 2000s, I observe lower point estimates and larger standard errors. These trends suggest that in later years the journal issue peer effects got weaker. Potentially, these trends are due to increase in online readership.

D.2 Conditionally Unconfounded Instrument

Unconfoundedness of the instrument requires that there are no unmeasured common causes between the instrument and the endogenous variable (follow-on citations), and between the instrument and the second stage outcomes of interest. I will discuss these assumptions and provide supporting evidence.

More prominent authors will tend to publish in more prestigious journals. The measurement of prominence through H-indexes varies across years. In addition, over time, some journals became more prestigious and influential compared to others. To account for these differences across journals and time, I condition all models on a strict set of fixed effects. Namely, I compare publications in different journal issues of the same journal and in the same year by including a set of journal-year fixed effects. Within a journal-year and given the first-in-first-out assignment process of manuscripts into journal issues, I claim that it is unlikely that confounders will drive the allocation of accepted manuscripts into specific journal issues. Exceptions to this assignment process are special issues and conference proceedings. Using indicators obtained from Web of Science and Dimensions data, I drop these cases from the sample.

Figure D1: First Stage Relation



Note: This figure presents a binned scatterplot of the relation between the logged sum of the top two researcher H-indexes (the instrument) and logged follow-on citations to the focal publication (endogenous variable). The values in the plot are fitted values after controlling for the logged H-index of the focal author, firm fixed effects and journal-year fixed effects.

To support the unconfoundedness assumption, Figure D3 presents a binned scatterplot of the relation between the instrument and the H-index of the top author of the focal paper. A corresponding linear regression reports a statistically insignificant slope estimate of 0.0093 (s.e. = 0.0099). In both the scatterplot and the regression estimates, there is no evidence that within a journal and year, more prominent authors jointly publish in specific journal issues. Nonetheless, in all model specifications I include the top focal H-index as an additional control.

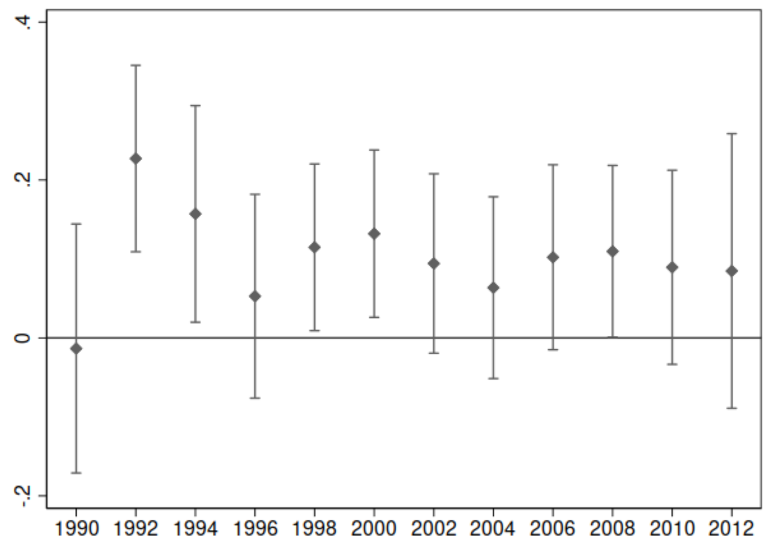
To further support the validity of the instrument, I perform a placebo test. In this test, I replace my instrument with a corresponding measure of top H-indexes from a randomly-picked journal issue within the same journal and year. In this test, the prominence of authors should not be relevant for follow-on citations. Figure D4 presents the test results. A corresponding linear regression reports a statistically insignificant slope estimate of 0.0143 (s.e. = 0.0158). According to the plot, there is no indication that prominence of authors from other journal issues drive attention (and therefore citations) to the focal publication. These results provide additional support for the mechanism that drives the relevance of the instrument.

Taken together, the results discussed above provide support for the assumption that the instrument is conditionally unconfounded.

D.3 Exclusion Restriction

The exclusion restriction posits that the prominence of other authors in the same journal issue as the focal paper affects outcomes only through their effects on external follow-on citations. A threat to this assumption could occur if, for example, information frictions within the firm limit internal awareness of the firm's own publications. In that case, the prominence of other authors in the same journal issue can drive more attention by the firm's own scientists, in the same way that it drives external attention to the

Figure D2: First Stage Time Trends



Note: This figure presents coefficient estimates for time trends of the first stage. The reported coefficients are of interactions between the instrument and 2-year indicators. The regression includes firm and journal-year fixed effects. Standard errors are clustered by firm.

focal publication. This possibility is highly unlikely, specifically when considering outcomes that directly relate to the focal authors (such as counts of their future publications and patents).

Another possibility is that the prominence of other authors will drive the firms' innovative outcomes through channels that are unaccounted for by the citation counts of follow-on research. For example, this can happen if some follow-on research does not cite the focal publication, but is found to be useful by the firm. While this is a possibility, it does not interfere with the general sense that follow-on academic activity can be beneficial for the firm.

D.4 Heterogeneity and the Average Causal Response

The literature on instrumental variables have long acknowledged the possibility of heterogeneity across the studied population (Angrist & Pischke, 2009). Under heterogeneity in observed and unobserved characteristics, instruments can only be used to estimate a local average treatment effect (LATE) instead of the average treatment effect (ATE). LATE refers to the average effect for a specific subset of the population, defined by their response to the instrumental variable. It may differ from the ATE, which is an estimate of the effect on the entire population. The difference between the LATE and the ATE depends on the degree of response heterogeneity and the strength of the instrumental variable.

The theoretical interpretation of the estimand is further complicated when the endogenous variable (the "treatment") and instrumental variable are continuous. Angrist and Pischke (2009) offer a generalization of the LATE framework to accommodate variable treatment intensity. The Average causal response (ACR) is a weighted average of the unit causal response, which in turn is the average difference in potential outcomes for compliers at different levels of treatment. When the treatment is fully continuous, IV estimation will recover the average derivative across the range of treatment values. When the instrument itself is continuous, the estimation will produce a weighted average of derivatives across the range of values of the instrument (Cornelissen et al., 2016).

Many of the coefficient estimates of the 2SLS regressions presented in this paper are larger than their

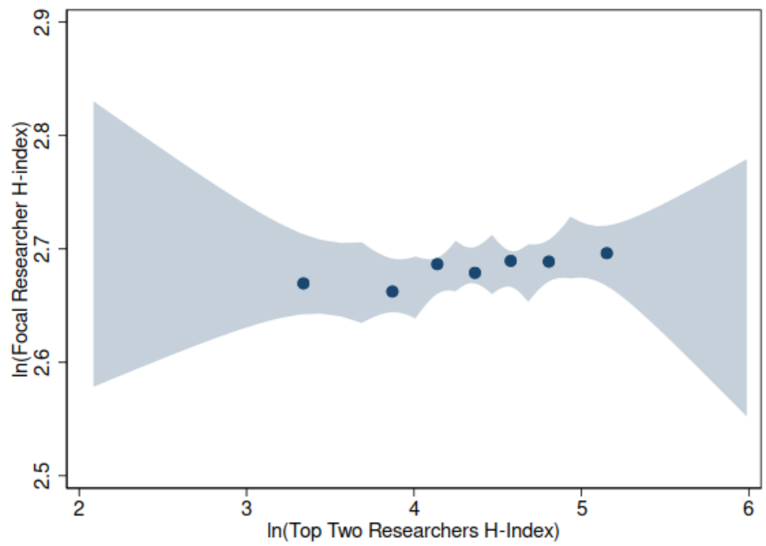
Table D1: First Stage Regressions

	ln(Follow-On)			
	OLS	OLS	OLS	OLS
	(1)	(2)	(3)	(4)
ln(Top One Researcher H-index)	0.075***			
	(0.016)			
ln(Top Two Researchers H-index)		0.097***		
		(0.018)		
ln(Top Two Researchers Pub. Count)			0.027**	
			(0.011)	
ln(Top Two Researchers Cit. Count)				0.046***
				(0.008)
ln(Focal H-Index)	0.280***	0.280***	0.280***	0.280***
	(0.010)	(0.010)	(0.010)	(0.010)
Firm FE	Yes	Yes	Yes	Yes
Journal-Year FE	Yes	Yes	Yes	Yes
Observations	164,495	164,495	164,495	164,495
Avg. DV	7.359	7.359	7.359	7.359
Adjusted R^2	0.589	0.589	0.589	0.589

This table presents estimation results for the first-stage relationship between the prominence of top researchers in the same journal issue as the focal publication and external follow-on research. In columns 1 and 2, prominence is measured using the authors' H-index in the year prior to publication. In column 3, prominence is measured as the previous publication count. In column 4, prominence is measured as the previous citation-weighted publication count.

Clustered (Firm) standard-errors in parentheses. Signif. Codes: ***: 0.01, **: 0.05, *: 0.1

Figure D3: Author H-index Correlation

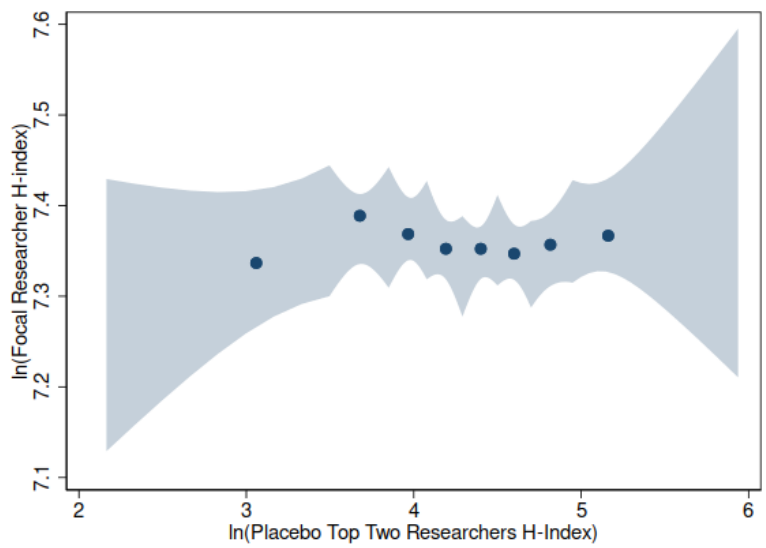


Note: This figure presents a binned scatterplot of the relation between the logged sum of the top two researcher H-indexes (the instrument) and logged H-index of the top focal author. The values in the plot are fitted values after controlling for firm and journal-year fixed effects. A corresponding linear regression reports a statistically insignificant slope estimate of 0.0093 (s.e. = 0.0099).

OLS counterparts. However, given the continuous nature of the instrument and endogenous variables, it is possible that these differences are due to the weighted nature of the ACR. Therefore, a direct comparison between the 2SLS and OLS estimates might be misleading and an analysis of the direction of bias is not trivial.

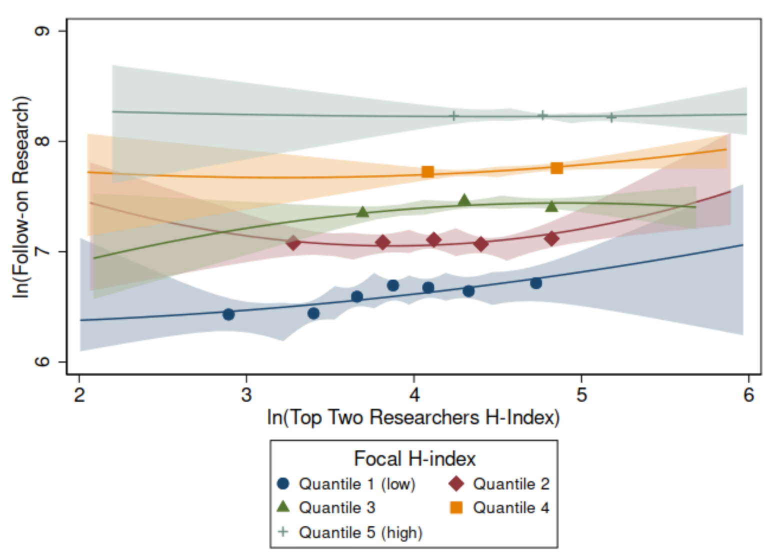
Further analysis of the first stage effects provides evidence for treatment heterogeneity across covariates. Figure D5 presents the first stage relation, across five quantiles of the focal authors' H-index. As expected, the effect is stronger for publications by less prominent authors. In addition, for these authors, the level of follow-on seems more strongly correlated with the probability of subsequent scientific publishing (Figure D6). While these relations seem not to hold for the case of patenting (Figure D7), the focal H-index is only one dimension of potential treatment heterogeneity.

Figure D4: First Stage Placebo Test



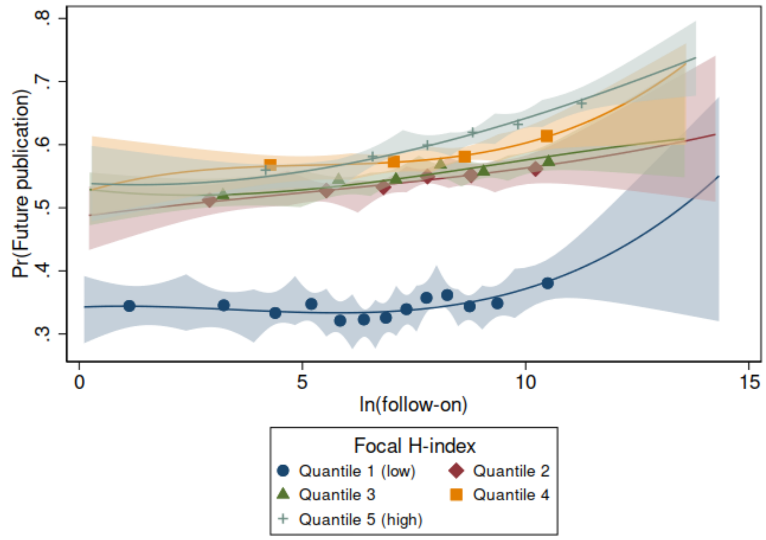
Note: This figure presents a binned scatterplot of the relation between the logged sum of top two researcher H-indexes from a random journal issue in the same journal-year (the placebo) and the logged follow-on citations to the focal publication (endogenous variable). The values in the plot are fitted values after controlling for firm and journal-year fixed effects. A corresponding linear regression reports a statistically insignificant slope estimate of 0.0143 (s.e. = 0.0158).

Figure D5: First Stage, by Focal H-index



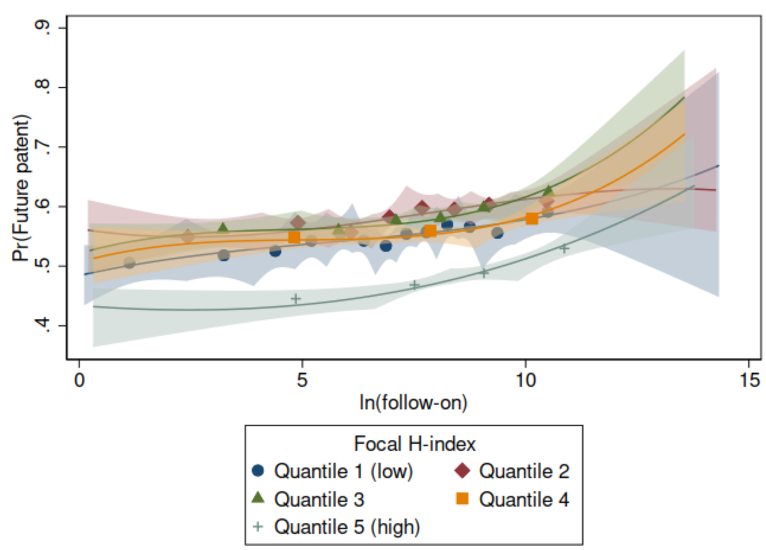
Note: This figure presents a binned scatterplot of the relation between the logged sum of the top two researcher H-indexes (the instrument) and logged follow-on citations to the focal publication (endogenous variable), across levels of the focal authors' H-index. The values in the plot are fitted values after controlling for the logged H-index of the focal author, firm fixed effects and journal-year fixed effects.

Figure D6: Follow-On Research and Subsequent Publications, by Focal H-index



Note: This figure presents a binned scatterplot of the relation between the logged follow-on citations and the probability of subsequent scientific publishing by the focal authors, across levels of the focal authors' H-index. The values in the plot are fitted values after controlling for the logged H-index of the focal author, firm fixed effects and journal-year fixed effects.

Figure D7: Follow-On Research and Subsequent Patenting, by Focal H-index



Note: This figure presents a binned scatterplot of the relation between the logged follow-on citations and the probability of subsequent patenting by the focal authors, across levels of the focal authors' H-index. The values in the plot are fitted values after controlling for the logged H-index of the focal author, firm fixed effects and journal-year fixed effects

D.5 Omitted Variable Bias in OLS

There are various sources of omitted variable bias in OLS estimates. A typical concern is the existence of an unobserved confounder (e.g., scientific quality). However, an additional source of bias, that is often overlooked, is a difference between the functional form of the data generating process and the observed variables. In such case, the direction of bias will depend on parameters of the model and can result in OLS estimates that are smaller than IV estimates, even when potential confounders would predict an upward bias of OLS. Consider the following data generating process:

$$\begin{aligned} X_1 &= Z_1 + \mu_1 \\ X_2 &= Z_2 + \mu_2 \\ X &= \delta X_1 + (1-\delta)X_2 \\ Y &= \alpha\delta X_1 + 0.5(1-\delta)X_2 + \mu_y \end{aligned}$$

The data includes two independent variables, X_1 and X_2. These variables are functions of instruments Z_1 and Z_2, respectively. Y is a linear combination of X_1, X_2, with δ defining the relative weights. The coefficient on $(1-\delta)X_2$ is set to 0.5 and the coefficient on δX_1 is α. Importantly, the researchers only observe X, Y and Z_2. Therefore, they estimate a 2-stage model as follows:

$$\begin{aligned} X &= \eta_0 + \eta_1 Z_2 + \xi \\ Y &= \beta_0 + \beta_1 X + \epsilon \end{aligned}$$

To clearly see the source of bias, consider the following rearrangement:

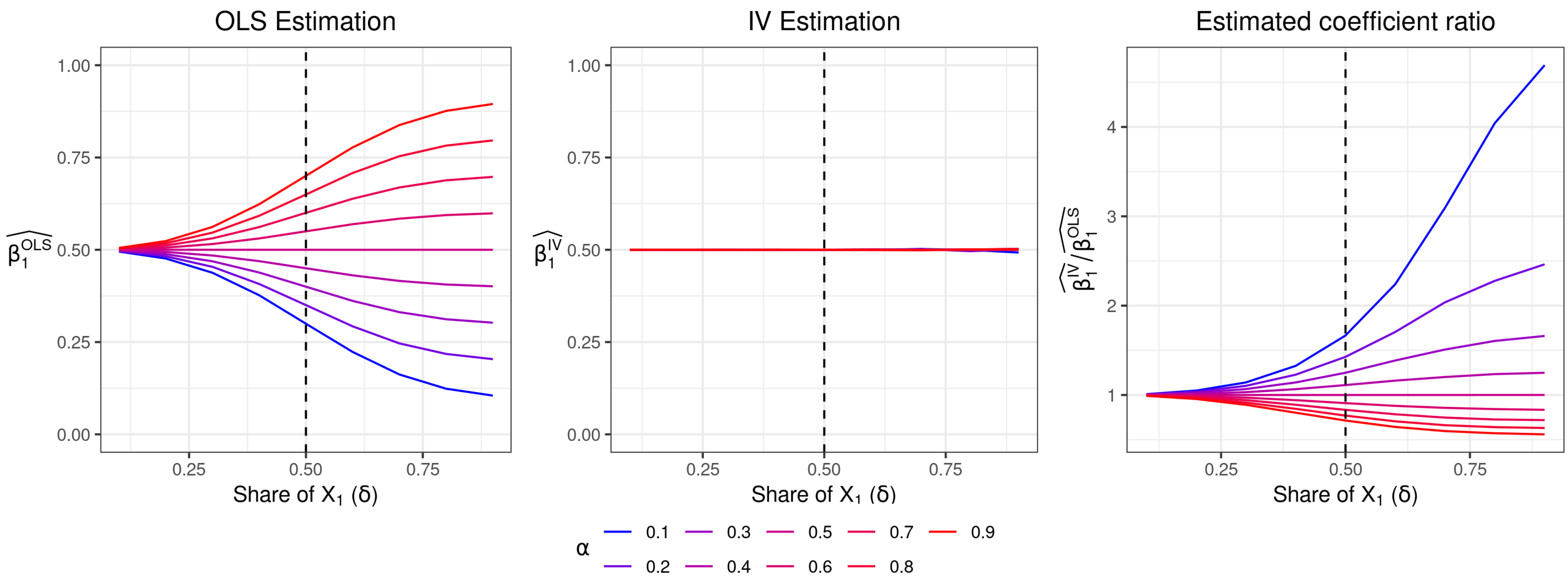
$$\begin{aligned} Y &= (\alpha\delta X_1 + 0.5(1-\delta)X_2 + \mu_y) + (0.5\delta X_1 - 0.5\delta X_1) \\ &= (\alpha - 0.5)\delta X_1 + 0.5\delta X_1 + 0.5(1-\delta)X_2 + \mu_y \\ &= 0.5X + U \end{aligned}$$

Where $U \equiv \mu_y + (\alpha - 0.5)\delta X_1$. The instrument Z_2 is uncorrelated with U, and therefore $\widehat{\beta}_1^{IV}$ would be unbiased regardless of the value of α. However, X itself is correlated with U through the joint dependence on X_1, and the direction of bias of $\widehat{\beta}_1^{OLS}$ will depend on the sign of $(\alpha - 0.5)$.

I provide an analysis of simulated data to explore this possibility. Results are presented in Figure D8. Clearly, when $\alpha = 0.5$ (i.e., the true coefficients on X_1 and X_2 are equal) then the OLS estimates are similar to the IV estimates. However, when $\alpha < 0.5$, then the OLS coefficient estimates are lower than the corresponding IV estimates, and the magnitude of difference increases with δ (the share of X_1 in Y). Alternatively, when $\alpha > 0.5$, the OLS coefficient estimates will be larger.

In the context of this paper, X is a number of citations for each publication and Z_2 is the observed instrument, however the instrument might influence only "marginal" citations, which are a part of X but unknown in size. Possibly, "marginal" and "core" citations have different effects on Y (note that the simulation uses normal distributions for simplicity). The implications of this analysis on the interpretation of results presented in the paper are that the functional form of the data generating process of citations can result in OLS coefficient estimates that are smaller than the IV estimates. If "marginal" citations have a stronger effect on the outcome, compared to "core" citations that are unaffected by the IV, then OLS coefficients will be smaller than the IV estimates. Note that this result can happen regardless of the direction of potential confounders and without treatment heterogeneity.

Figure D8: OLS and IV Coefficient Estimates, Simulation Results



Note: This figure presents simulated results comparing OLS and IV coefficient estimates. The parameters used are as follows: $Z_1 \sim \mathcal{N}(0, 25)$, $Z_2 \sim \mathcal{N}(0, 25)$, $X_1 = 10Z_1 + \mu_1$, $X_2 = 10Z_2 + \mu_2$. All noise variables (μ_1, μ_2, μ_y) are distributed $\mathcal{N}(0, 100)$. Sample size is 50000.