Patent Class Contrast and Patent Citation Impact

Balázs Kovács
Yale University
School of Management
165 Whitney Ave
06520 CT, USA
balazs.kovacs@yale.edu

Gianluca Carnabuci
ESMT Berlin
Schlossplatz 1
10178 Berlin, Germany
gianluca.carnabuci@esmt.org

Filippo Carlo Wezel
Universita Svizzera italiana
Via Buffi 2
Lugano, Switzerland
wezel@usi.ch

Forthcoming in *Strategic Management Journal*

Abstract

Whereas prior innovation and strategy literature studied how attentional and search dynamics influence the creation of inventions, we examine how these same processes affect the impact of inventions *after* their creation. We theorize that inventions classified in "high-contrast" technological categories garner more attention by potential users and, hence, accrue more citations than otherwise-equivalent inventions classified in "low-contrast" categories. We test this hypothesis via three studies. First, we estimate citation-count models among all USPTO patents granted between 1975 and 2010. Second, we conduct a "twin patents" test comparing inventions patented both at the USPTO and at the EPO. Third, we examine minute-by-minute search logs from a sample of USPTO examiners. These studies support our hypothesis and extend current understandings of attentional and search dynamics in the innovation process.

INTRODUCTION

Building on the concept of bounded rationality (March & Simon, 1958; Nelson & Winter, 1982), a considerable body of literature has examined how processes of selective attention allocation and information search influence the creation of innovative ideas, products, and technologies (for a review, see Eggers & Kaplan, 2013). The point of departure of this literature is that inventions are created by recombining existing knowledge (Schumpeter, 1934). Therefore, where actors focus their limited attention as they search for the knowledge inputs affects the quality of their inventive output (Ghosh et al., 2014; Dahlander et al., 2016). For example, Martin and Mitchell (1998) found that most firms have a narrow focus of attention and search for knowledge inputs locally, i.e., close to what they already know; however, engaging in distant search helps firms develop products that are more unique and innovative. Similarly, Ferguson and Carnabuci (2017) argue that searching for knowledge inputs across technological boundaries improves an inventor's chances of identifying superior technological solutions.

Whereas this line of inquiry leveraged theories of attention and information search to illuminate the *creation* of inventions, it did not examine whether processes of attention allocation and information search also affect the impact of an invention *after* it has been created. Unquestionably, inventions of higher technical quality are likely to have greater impact (Moser et al., 2018), i.e., to be more widely used and "built upon as technology continues to evolve" (Rosenkopf & Nerkar 2001: 291, Fleming, 2001). However, focusing on what happens after an invention has been created is important because the impact of an invention does not only depend on its inherent technical quality but also on whether potential users become aware of its existence and perceive it to be useful as *they* search for knowledge inputs to build upon (see Mokyr, 2002; Furman & Stern, 2010; Bikard, 2018). Because inventions burgeon at a too fast

rate for anyone to comprehensively search them all, many technically valuable inventions may pass unnoticed and become dead-ends (Fleming, 2001), while technically subpar inventions may accrue significant attention and impact (Greve & Seidel, 2015). Consistent with this view, the present paper shifts the focus of analysis from the attentional and search dynamics that shape the process by which inventions are created, to those that guide the attention of potential users once an invention has been created and is ready for use.

Similar to prior studies, we focus on a particularly important inventive output – patents – and measure their impact by tracing the citations they receive from future patents (e.g., Rosenkopf & Almeida, 2003). Patents provide an ideal context to investigate our research question because patent users[1] face considerable information overload, meaning that the volume of available information exceeds their cognitive capacity (Lemley, 2001; Kuhn, 2010), and must cope with it by allocating their attention selectively. Prior studies show that patent users' search for knowledge inputs is shaped by informational cues related to the status of a patent holder (Podolny & Stuart, 1995), and to the institutional (Bikard, 2018) and geographical origins of a patent (Bikard & Marx, 2020). This literature suggests that examining where patent users are more likely to focus their limited attention is helpful to understand why, net of their technical value, some patents accrue more citations than others.

[1] Throughout the paper, we will use the expression "patent user" to indicate anyone who is actively involved in the patenting process, whether as a patent applicant or as a patent examiner. We are aware that the roles of patent applicant and patent examiner are distinct and that, in fact, even more fine grained distinctions (e.g., between inventors and patent attorneys) may be relevant when examining the patent citation process. We are also aware that there are differences across patent offices, e.g., the EPO and the USPTO. As we explain in more detail later in the paper and in the appendix, we devised several approaches to account for these differences; in fact, we leverage some of these differences in our empirical tests to further probe our argument. From a theoretical perspective, however, our focus is on any expert who actively engages in searching for relevant prior art either during the inventive process (e.g., when an inventor draws ideas for a new process or device by scouting existing patents) or the patent application process (e.g., when a patent applicant or a patent examiner searches for relevant patents to be listed in a patent application document).

We contribute to this line of inquiry by arguing that patents classified in *high-contrast* technology classes are more likely to receive attention, and hence to be cited, compared to otherwise-equivalent patents classified in *low-contrast* classes. Our arguments build on and extend the work of Zuckerman (1999), who proposed that the industry categories used to classify equities effectively delimit the "consideration set" on which stock analysts focus their attention. Similarly, we argue that patent classes, by partitioning the stock of existing patents into distinct technological categories, enable patent users to narrow down the consideration set on which they focus their limited attention when searching for relevant prior art. We recognize that some patent classes have high category contrast, i.e., they demarcate a distinctive and well-defined technological category, while others have low contrast, meaning that the boundaries separating them from neighboring categories are blurred (Hannan, 2010; Carnabuci et al., 2015). Extant categorization research suggests that high-contrast categories are more salient and more informative than low-contrast ones (e.g., Hsu, 2006a; Negro et al., 2015). As a result, the information classified in high-contrast categories receives more attention, whereas that in *low-contrast* categories is more likely to pass unnoticed (Rosch, 1975; Murphy, 2004; Hannan, 2010). Building on this argument, we posit that patent users are more likely to focus their limited attention on high-contrast than on low-contrast patent classes when searching for relevant prior knowledge. Therefore, holding other things constant, the number of citations a patent receives should be greater if it is classified in a high-contrast patent class than in a low-contrast one.

Testing our theory poses a methodological challenge. Our goal is to isolate the effect of category contrast on patent citations from possible confounding factors, including patents' unobservable technical quality. We address this challenge by conducting three related studies. In Study 1, we establish the plausibility of our hypothesis by estimating a set of panel count models

with patent and year fixed effects, using the entire record of utility patents granted by the United States Patent and Trademark Office (USPTO) between 1975 and 2010. We find a strong and robust effect of patent class contrast on patent citations. To more conclusively isolate the causal effect of category contrast on patent citation, in Study 2 we conduct a "patent twins" case-control design (Bikard, 2020) leveraging cases of identical inventions patented in two distinct patent offices – the USPTO and the European Patent Office (EPO). Because these patent twins refer to the exact same invention, which is subject to parallel assignments to multiple classification systems, this identification strategy enables us to tease out the effect of patent class contrast on patent citations while perfectly controlling for the underlying technological quality and other unobservable time-varying factors. We find that patent class contrast continues to exhibit a sizeable and robust positive effect on patent citations even when matching patent twins in a case-control design.

Study 1 and Study 2 provide compelling empirical support for the hypothesis that inventions classified in higher-contrast classes receive more citations, but they do not enable us to directly observe the attention allocation and search mechanisms that we postulate to drive this effect. Whereas micro-level search behavior data is exceedingly rare, it does exist for at least one kind of patent users – patent examiners – enabling us to probe our theorized mechanisms more directly. Thus, in Study 3, we analyze a unique, minute-by-minute data set obtained from the search log records of 610,764 queries conducted by USPTO patent examiners searching for relevant prior art for 17,373 patent applications. We find that the search behaviors of this subset of patent users confirm the mechanism postulated by our theory. First, patent examiners are significantly more likely to focus their prior art searches on high-contrast than on low-contrast patent classes. Second, the higher the contrast of a patent class, the more likely are patent

examiners to delimit the scope of their prior art search to patents just inside that class. Third, when patent examiners search for prior art in high-contrast classes, their searches are more efficient, i.e., they add more citations in less time.

This study contributes to the literature by extending current understandings of attentional and search dynamics in the innovation process. Prior innovation and strategy literature examined how attentional and search processes influence the creation of inventions, finding that those who search for knowledge inputs across technological boundaries are more likely to identify superior technological solutions. Shifting the focus analysis from the creators to the users of inventions, conversely, we highlight how inventions classified within clear-cut, high-contrast technological boundaries tend to garner more attention and, therefore, exert greater impact on future technological developments. By demonstrating that a seemingly neutral and inconsequential classification decision may systematically influence the future impact of an invention, our study contributes new insight into the growing body of literature that examines what drives a patent's future use, and hence value as a knowledge asset, over and beyond a patent's inherent technical quality (Murray & O'Mahony, 2007; Ferguson & Carnabuci, 2017; Bikard, 2018; Polidoro, 2020). On a more programmatic level, our study underlines the importance of categorization processes in strategy and innovation research (e.g., Pontikes, 2018).

THEORETICAL BACKGROUND

The number of citations a patent receives ultimately depends on whether patent users find it as they search for relevant prior art. How, then, do patent users search for relevant prior art across the overwhelming stock of patented knowledge? Since March and Simon (1958), we know that people cope with information overload by engaging in heuristic search for satisfactory rather

than optimal solutions. Research in strategy (Stuart & Podolny, 1996; Katila & Ahuja 2002; Rosenkopf & Almeida, 2003), innovation studies (Fleming, 2001; Arthur, 1989), and psychology (Newell & Simon, 1972; Gigerenzer et al., 1999) has consistently shown that people are boundedly rational and manage information overload by taking cognitive shortcuts, or "heuristics" (Kahneman et al., 1982), which focus their attention on a selected subset of the available information. The extent to which people rely on cognitive shortcuts may vary depending on how much prior domain knowledge a person has; however, as Simon (1971, 1986) and subsequent studies have compellingly shown, even experts cope with information overload through a selective attention allocation process (for a comprehensive review, see Campitelli & Gobet, 2010). For example, Bikard (2018) examined how R&D scientists identify relevant academic research among the stock of articles published in academic journals. Leveraging a "paper twin" analysis, he compared independent publications of the very same discovery, such that the likelihood of being cited should be exactly equivalent for both paper and "twin." He found that R&D scientists use the papers' institutional origin as a cognitive shortcut when choosing where to allocate their scarce attention. This, in turn, leads them to selectively focus on certain publications while systematically disregarding others that are equally relevant.

Categories play a particularly important role in focusing people's selective attention during information processing tasks (Hannan et al., 2007; Zuckerman, 1999; Ruef & Patterson, 2009). By breaking down the stock of available information into subsets of related information, categories reduce the "consideration set" to be searched by users and orient their focus (Posner & Petersen, 1990; see also Helfat & Peteraf, 2015). Research in a broad variety of settings including the stock market (Zuckerman, 1999), films (Hsu, 2006a, 2006b), music (Montauti & Wezel, 2016), and restaurants (Kovács & Hannan, 2010) has shown that categories shape where

people allocate their attention, investments, and resources. More recently, this line of inquiry expanded to examine the role of a particular type of category – patent classes – finding that they affect firms' technological entry decisions (Carnabuci et al., 2015), the ways in which the portfolio of investments of technology startups is evaluated by venture capitalists (Wry & Lounsbury, 2013; Wry et al. 2014), and the likelihood of patent rejections (Ferguson & Carnabuci, 2017). As we articulate in the next section, we expect patent classes to also affect the likelihood that a patent will be cited by future patents.

Hypothesis development

Just like scholars constantly search for relevant academic literature, a primary task of every patent user – be it an inventor, patent examiner or patent attorney – is to identify relevant prior art. At the time of this writing, there are over 11 million patents in the USPTO alone. In the absence of a patent classification system, finding relevant prior art across the stock of patented knowledge would be akin to finding a needle in the haystack. Patent classes sort related patents into technological categories, and in so doing they greatly simplify the search for relevant prior art, making it pragmatically manageable (Kuhn, 2010). Building on Zuckerman's (1999) argument that categories narrow down users' "consideration sets," we argue that patent classes work as an "information infrastructure" (Bowker & Starr, 2000) that channels the attention of patent users and selectively focus their search for prior art[2].

[2] This purpose of the patent classification system is also explicitly stated in USPTO documents. For example, the document titled "Overview of the U.S. Patent Classification System (USPC)" states that "The USPC serves (…) to facilitate the efficient retrieval of related technical documents" (see page I-2, https://www.uspto.gov/sites/default/files/patents/resources/classification/overview.pdf)

Extant research posits that the degree to which a category channels users' attention varies with its *contrast*, defined as the extent to which a category contains unique and distinctive information that sets it apart from neighboring categories (Hsu, 2006a; Negro et al., 2015; Kovács & Hannan, 2010; Hannan et al., 2019). In the context of our study, specifically, a patent class exhibits high category contrast if most of the patents categorized therein belong *only* to that class. Conversely, a patent class is low-contrast if its categorical boundaries are unclear, that is, many of the patents categorized therein are also categorized in other classes. Building on categorization theory (e.g., Hannan et al., 2007), we propose that high-contrast classes are more likely to delimit patent users' consideration sets because they are both more *salient* and more *informative* than low-contrast ones. They are more salient because, being clearly different from neighboring classes, they "stand out from the rest" (Hannan, 2010; Hannan et al. 2019). They are more informative because they effectively bracket most of the potentially relevant prior art and, therefore, they provide a useful indication of the boundaries within which patent users should focus their search. Low-contrast classes, on the other hand, do not univocally demarcate the boundaries within which patent users should search for relevant prior art because they comprise a larger share of patents that are concurrently categorized in one or more neighboring classes. Hence, they are both a less salient informational cue and a less useful heuristic for focusing the scope of patent users' searches.

These arguments suggest two reasons why, holding other things constant, patents classified in high-contrast classes are more likely to be identified as relevant prior art and, hence, accrue more citations than patents in low-contrast classes. First, patent users are more likely to start their prior art search in high-contrast patent classes rather than on low-contrast ones because the former, being clearly distinct from neighboring classes, provide a clear indication of the

"consideration set" within which prior art should be searched. Conversely, low-contrast classes do not provide a clear indication of where patent examiners should focus their attention because they are less distinguishable from other potentially relevant classes. In fact, in cases of extremely low category contrast, classes provide no heuristic value at all, similar to a situation where there is no classification. Second, when searching for prior art within a high-contrast class, patent users are more likely to focus the scope of their search within the boundaries of the class itself, rather than dispersing their attention across multiple classes. Such narrow focus is cognitively efficient (Castiello & Umiltà, 1990) and, therefore, increases the likelihood that relevant patents classified therein will be identified within the patent users' limited time budget. Conversely, when patent class contrast is low, patent users must spread their attention more broadly and search for potentially relevant citations across multiple classes. Such lack of attentional focus does not only reduce the time[3] patent users can devote to each class searched, but also the efficiency, and hence the returns, of the search process. Evidence indicates that dispersing one's attention across multiple classes is cognitively inefficient and drastically diminishes the chances of finding relevant information (e.g., see Cowan, 2016). Hence, when patent users search for prior art in a low-contrast class, the likelihood that relevant prior art will pass unnoticed – and hence will not be cited – increases sharply.

To embed these theoretical arguments within the reality of the patenting process, it may be useful to consider the following thought experiment. Imagine a patent examiner tasked with finding relevant prior art for a patent application[4]. In the spirit of a counterfactual analysis, let us

[3] Research has documented that examiners typically have one or at most a few hours per application for prior art search (Frakes & Wasserman, 2017).

[4] The thought experiment focuses on patent examiners, but it could be readily applied to any patent user searching for relevant prior art, e.g., an inventor trying to map the body of existing knowledge around a new device she has been working on.

compare two identical scenarios where the only difference is in the category within which patents are classified. Because of this difference, the category contrast of the patent class in question is high in one scenario and low in the other. For simplicity, let us use the shorthand Class *x* to denote the high-contrast scenario and Class *y* to denote the low-contrast one. While these are hypothetical counterfactual scenarios, we note that the statistics used to describe represent the 10^{th} and 90^{th} percentiles of the observed distributions.

Scenario x: With a category contrast score of 0.89, Class *x* is one of the highest-contrast classes and, therefore, it is clearly distinguishable from neighboring classes. For example, 78% of the patents classified in *x* are classified only within *x*. Furthermore, 73% of the prior art citations from patents classified in *x* reference earlier patents within the class itself. That is, class *x* contains most of the relevant prior art for patents therein.

Scenario y: With a category contrast score of 0.65, class *y* is one of the lowest-contrast classes. Class *y* is hardly distinguishable from neighboring classes *k*, *j* and *z* because many patents classified in *y* (64%) are concurrently classified in *k*, *j* and *z*. Furthermore, a large share of the relevant prior art cited by patents classified in *y* comes from other classes, with *y* bracketing on average only 38% of the relevant prior art.

How would our hypothetical examiner behave in these counterfactual scenarios? Our argument is that, because *x* is clearly distinguishable from its neighboring classes, it is more salient and therefore likely to catch the examiner's attention as she sets out to define her "consideration set" and start her search for prior art. Furthermore, *x* is informative: if our examiner decides to begin searching within *x*, she is likely to find that most relevant prior art is classified therein. Therefore, she will likely deepen her search within the class rather than spreading her attention across other classes. Because focusing on a single class increases the

cognitive efficiency of the search process, chances that our examiner will be able to identify relevant prior art given her limited time budget are higher. For these reasons, any relevant patent that is classified in *x* stands a comparatively high chance of being cited.

The situation is markedly different in the counterfactual scenario where the relevant class is *y*. First, because *y* has low category contrast, it is not clear where its boundaries stop and where the boundaries of neighboring classes *k*, *j* and *z* begin. Hence, the examiner would have to first ask herself where to begin her prior art search. Class *y* might be a candidate, but it certainly does not stand out as the only one. Because most of the patents in *y* are concurrently classified in *k*, *j* and *z*, these other classes are good candidates, too. Thus, compared to class *x*, class *y* provides a feebler informational cue of where our examiner should delimit her "consideration set" at the outset of her prior art search. Second, even if our examiner did decide to initially focus her attention on *y*, she would soon realize that doing so is not particularly helpful in terms of delimiting the boundaries of her search scope because much of the relevant prior art resides outside of *y*. As a result, she would have to spread her search thinly across multiple patent classes, which reduces the cognitive efficiency of the search process and, therefore, increases the likelihood that our examiner will miss potentially relevant prior art.

In synthesis, we posit two reasons why, *ceteris paribus*, patents classified in high-contrast classes are likely to accrue more citations than patents in low-contrast classes. First, patent users are more likely to start their search and focus their attention on high-contrast than low-contrast patent classes. Second, patent users are cognitively more efficient when searching for prior art in high-contrast rather than low-contrast classes and, hence, they are more likely to identify relevant prior art in the former than in the latter. These arguments lead to the central hypothesis of the paper.

Hypothesis: The higher the contrast of a patent class, the higher the number of citations the patents categorized therein will receive.

OVERVIEW OF THE THREE STUDIES

Because patents are not assigned to patent classes randomly, demonstrating the effect of patent class contrast on patent citation poses an identification challenge. Most notably, it is possible that the patents classified in high-contrast classes are qualitatively different (e.g., better) than those in low-contrast classes. To address this challenge, we conducted three related studies. Table 1 provides an overview of the three studies and how they relate to each other. We discuss each study's logic, sampling strategy and methodology in the next sections.

---------- Insert Table 1 Here ----------

To gain a better insight into the classification and patent citing processes, we conducted a number of field interviews. Specifically, we interviewed three current USPTO patent examiners, a former head of the patent classification office at the USPTO, and a senior patent attorney who has worked with Silicon Valley startups and large hardware/tech firms. The interviews were conducted via Skype and lasted around 45-60 minutes each. The interviews followed a semi-structured format, in which we asked questions about (i) the experience of the interviewee with patenting, (ii) the patenting process and prior art search at the USPTO and (iii) the potential for strategic behaviors from the side of applicants. In addition to the interviews, the first author

completed the first step of the patent examiner training at the USPTO, which provided further understanding of the setting and of the interpretation of the examiner search data used in Study 3.

STUDY 1

In Study 1, we model patent impact as a patent's yearly citation count and estimate a set of panel count models with patent and year fixed effects. We note that this is a statistically more conservative approach relative to most existing studies of patent citation. Most prior studies primarily controlled for patent heterogeneity by measuring observable characteristics, such as number of prior art citations, technological breadth, component familiarity, combination familiarity, scientific references, patent's primary industry, number of inventors, or age, status and size of the patent assignee (see Fleming, 2001; Fleming & Sorenson, 2004; Carnabuci et al., 2015 for details on these variables). By estimating a patent-level fixed effects model, we remove any observable and unobservable differences across patents, including all the characteristics controlled for by prior studies (Alcácer & Gittelman, 2006). Furthermore, by including year fixed effects, we control for possible time trends and time-vary shocks that may affect all patents simultaneously, such as economic cycles or changes in patenting policy and regulations of the patent office. For example, the USPTO changed patent examiners' remuneration policy and moved to a new headquarters in 2005. Such changes are hard to trace comprehensively but can be effectively controlled for by including year fixed effects.

Data and sample

In Study 1, we analyze all utility patents granted by the USPTO between 1975 and 2010. The year 1975 was chosen as a starting point because since that year the records are made

available in digitized format (Jaffe & Trajtenberg, 2002). Our data end with the patents granted before or in 2010 because in 2011 the USPTO started to introduce a new classification system, CPC, which went into effect in 2013. To provide a clean test of the effect of patent class contrast, we restrict our sample to patents that were not reclassified until the end of our observation window, resulting in a final sample of 3,260,030 patents. In Study 1, we utilize the patent dataset made available by the USPTO at Patentsview.org.

Measures

Patent Citations. This is our primary dependent variable in Studies 1 and 2. Following a consolidated practice (Hall et al., 2005), we calculate patent impact yearly, based on the citations it receives from patents applied for in a given year. In the USPTO, patent citations can be added by either patent applicants or patent examiners. In Study 1, we measure patent impact as "total citations received" regardless of who inserts the citations. This is for two main reasons. First, this is the dependent variable on which most prior patent impact studies are based, which increases the comparability of our findings to previous work. Second, the USPTO did not record, for most of our observation period (1975-2000), who added the citations. As we will show later, the results presented are robust to alternative time windows and specifications, such as taking the 2001-2010 subsample or focusing on examiner added citations only.

Patent Class Contrast. Testing our hypothesis requires a measure of the contrast of each patent class. To this end, we build on the conceptualization of Hannan and colleagues (2007) and operationalize contrast through the measure developed by Carnabuci and colleagues (2015)[5]. The measure is constructed in two steps. *Step 1* leverages the fact that, whereas in the USPTO each

[5] In the appendix, we report analyses based on different measurement approaches and find that our results remain highly consistent irrespective of measurement approach.

patent can only be classified in one primary class, the technological subclasses assigned to the patent may belong either to the primary class itself or to other classes. This design feature enables us to build a patent-level "recombinativeness" score that measures the proportion of technological subclasses assigned to a patent that do not belong to the primary class within which the patent is classified (Carnabuci et al., 2015).[6] The recombinativeness measure ranges from 1/n to 1, where n denotes the number of classes the patent is assigned to. In *step 2*, we calculate the contrast of each patent class as one minus the average recombinativeness of the patents assigned to the focal class in a given year. Thus, a patent class has a category contrast of one when all the patents classified therein are exclusively classified in that class and no other. However, the more the patents classified in a class combine technological subclasses belonging to other primary classes, the closer class contrast approaches zero.[7]

As new patents are granted every year, this measure is updated yearly. We measure contrast in terms of flow rather than stock to capture the possibility that a patent that belongs to a low contrast category today might actually gain visibility over time as the class gains contrast (and vice versa). Nonetheless, while patent class contrast does vary over time, the 5-year autocorrelation value of 0.62 indicates its relative stability in the short- and medium-term. Table 2 provides illustrations of patent classes by listing the primary classes with the 10 highest

[6] For example, US patent 4,836,431 is granted for a "Semi-automatic loading paper feed tractor" that is "provided with an endless belt which travels in a triangular path and a drive shaft." In laypeople's terms, this is the belt mechanism that moves along the printing paper in the matrix printer (the printer with perforated holes, popular in the 1980s and 1990s, still used at some airports). This innovation draws on two bodies of knowledge: printers and belt-mechanisms. Reflecting the underlying technological recombination, this patent is classified as 226/74 "Advancing material of indeterminate length" and 400/616 "Typewriters", with the primary classification being 226. Because one (226) out of the two classes (226 and 400) is the same as the patent's primary class, the patent's recombinativeness value is 0.5. An example for a non-recombinative patent is US patent 6,652,318 "Cross-talk canceling technique for high speed electrical connectors," which is an innovation that provides a new configuration for electrical connectors and is classified only in primary class 439 ("Electrical connectors"). Therefore, this patent's recombinativeness score is 0.

[7] We also estimated our models with alternative specifications of the contrast variable, see the results in Tables A1-A3 in the Appendix.

average contrast class (e.g., "Static information storage and retrieval" or "Electrical Connectors") to the 10 lowest average contrast class (e.g., "Synthetic Resins or Natural Rubbers" or "Concentrating evaporators"). Figure 1 shows how category contrast changes over time for three randomly selected patent classes.

Control variables. In addition to incorporating patent-level and year fixed effects, we control for two time-varying changes at the patent-class level that prior studies have identified as affecting the citations received by a patent. Specifically, as patent classes vary in popularity and R&D investments, we control for class-specific technological fertility and growth rates (Fleming, 2001; Fleming & Sorenson, 2004; Carnabuci & Bruggeman, 2009). Following common practice in the innovation literature, we measure *Technological fertility* as the mean number of citations received by the patents that are classified in the primary class in the given year. Additionally, we control for the yearly number of patents granted in each class (*Class size)* as previous studies found that citation growth rates are related to patent class size (Carnabuci, 2013; Lafond & Kim, 2017). Note that because the models presented include patent fixed effects, we do not need to (and may not) include additional controls for factors that are constant for a given patent, such as the count of inventors, the count of claims, its degree of recombinativeness, the count of scientific references, or the nationality and size of the assignee.

---------- Insert Figure 1 and Tables 2 and 3 Here ----------

Estimation method

We estimate variations of the following equation:

EQ (1): $\theta_{i,t+1} = f(\varepsilon_{i,t+1}; \alpha_{t+1} + \partial_i + \gamma_{i,t} + \beta_{i,t,j})$,

where the index *i* and *t* identifies the patent and the year, $\theta_{i,t}$ denotes the dependent variable, the count of citations patent *i* receives in year *t*; α_t is a year fixed effect, ∂_i is a patent fixed effect, $\gamma_{i,t}$ is the contrast of the patent's primary class. $\beta_{i,t,j}$ captures relevant control variables, such as the fertility of patent class and the size of the patent class. All independent variables are lagged by one year. Because these are patent fixed-effects models, and because each patent belongs to a single primary patent class, these models leverage the over-time variance in patent class contrast. We report standard errors clustered on patents, but we note that the results are robust to other standard error calculations as well, such as robust standard errors or clustering on year and patent classes (Wooldridge, 2002; Mehta et al., 2010).

Results

Table 3a reports descriptive statistics and pairwise correlations between the main variables. Of particular note here is the positive correlation between patent class contrast and future citations. This provides a *prima facie* evidence of a positive relationship between class contrast and citation counts[8]. Table 4 reports results obtained from multiple specifications of equation (1) above. Model 1 is a log-linear specification (where the DV is the logged value of one plus the count of forward citations made to the patent). Model 2 is a Poisson model, while Model 3 is a negative binomial model.

We begin by pointing out that about 43% of the overall variance is explained by the patent fixed effects. This is likely to reflect underlying differences in patent quality, in line with the intuition that technically better inventions are more cited. Controlling for this heterogeneity is exactly the reason why we included patent fixed effects. Net of such patent-level differences, the

[8] We thank an anonymous reviewer for pointing this out.

estimated effect of patent class contrast is positive and statistically significant across all three models. In terms of effect size, the results show that one standard deviation increase in the patent's class contrast leads to between 1.7% and 2.6% increase in citation in the following year (based on the estimates from Table 4, Models 4 and 6). It is hard to directly compare this effect size to those reported by prior studies because most prior studies did not include patent fixed effects. Nevertheless, we notice that the effect size of class contrast is likely to be larger than that of, for example, combination familiarity (0.6%, see Fleming, 2001) and cumulative combination usage (0.1%, see Fleming, 2001), while it is likely smaller than the effect of using scientific literature in the inventive process (3.8%, see Fleming & Sorenson, 2004) and knowledge breadth (5.6%, see Fleming & Sorenson, 2004).

As robustness checks, we conducted analyses with alternative specifications of patent class contrast and found that the effect of this variable remains consistent across multiple specifications, such as when using a contrast value smoothened in different ways or when resorting to alternative measures of patent class contrast (see Table A1 in the Appendix). Taken together, these results provide robust evidence for our hypothesis and show that patents classified in higher-contrast patent classes tend to receive more citations.

---------- Insert Table 3a and 4 Here ----------

STUDY 2

The patent and year fixed-effects specifications of Study 1 control for time-invariant patent-level heterogeneity as well as for time-varying factors that affect all patents similarly. However, citation patterns may reflect life-cycle differences across technologies and patent

types. For example, the citation trajectories of "Electrical" and "Manufacturing" patents peak two years earlier than that of "Drug" patents (Mehta et al, 2010). Simple patent-fixed effects do not pick up such effects, nor do they pick up more complex interaction effects between time-invariant patent characteristics (e.g., number of inventors, number of classes, assignee size etc.) and technological trajectories. To overcome this problem and more conclusively identify the causal effect of class contrast, in Study 2 we rely on a twin-design (Bikard, 2020). Specifically, we use a "patent twins" case-control design in which we exploit cases of identical patents that are granted both at the USPTO and at the EPO and, therefore, are subject to parallel assignments into two classification systems. Because these patent twins refer to the exact same underlying technological invention, this identification strategy enables us to tease out the effect of class contrast on patent citation while perfectly controlling for the quality and other unobservable traits of a given patent, as well as their interactions with technological trajectories.

Data, sample and additional measures

Like Study 1, Study 2 analyzes patent citations in a yearly-panel format. However, the sample in Study 2 differs from Study 1's sample in three ways. First, we restrict our attention to patents that were submitted simultaneously to the USPTO and EPO and were granted at both patent offices. This subset allows us to build a control group of "twin patents" that we can match with our focal patents. To identify the sample, we rely on the patent family information provided in the Triadic Patent Database (Dernis & Khan, 2004)[9]. Since patent families sometimes

[9] Triadic patents are not a random subset of all patents at the USPTO. As Kovács (2017) demonstrates, USPTO patent applications that are also submitted to EPO, in comparison with USPTO patent applications that not submitted to the EPO, (i) are assigned to more subclasses, (ii) are less focused, (iii) have more independent claims, (iv) have more inventors, (v) are less likely to be a small entity, (vi) are less likely to be US-domestic, (vii) more likely to be maintained, and (viii) receive fewer citations if accepted. The lack of representativeness of this subset of

consolidate multiple related patents, in order to ensure exact matches, we only analyze patent pairs that (1) have a 1-to-1 match between the USPTO and EPO version, (2) have the exact same title in both the USPTO and the EPO, and (3) were not reclassified either in the USPTO or at the EPO. Second, unlike at the USPTO, at the EPO citations can only be added by patent examiners whereas patent applicants are not allowed to add citations. To maximize comparability between the two offices, when counting the number of citations received by a patent in the USPTO, we focus on examiner-added citations only and disregard applicant-added citations.[10] Focusing on examiner-added citations also allows us to get a cleaner test of our proposed mechanism because examiners' primary task is to find and cite all relevant prior art (USPTO Manual of Patent Examination, 2015), while applicant-added citations may be biased by multiple strategic motivations (Lampe, 2012). Third, because data on examiner-added citations at the USPTO are only available from 2001, we restricted our sample to patents granted by both the USPTO and the EPO between 2000 and 2010. The one-year difference is because the variables are lagged.

A total of 67,389 patent-twins satisfy the three sample inclusion criteria. These patent-twins translate into 621,925 patent-twin/year observations. The panel, as in Study 1, is unbalanced.

patents, however, is less of a worry here, as the goal of these tests is to demonstrate the marginal effect of changes in category contrast among a set of identical-invention pairs.

We also acknowledge that the USPTO and the EPO differ not only in their classification system, but also in other aspects such as the overall number of prior-art citations. While these are relevant differences, what matters for the purpose of our test is that citation counts capture impact in both settings – a claim that is undisputed in the patenting literature. By modeling patent-twin fixed effects (i.e., technology fixed effects), our tests adequately control for differences in citation (and other) practices as long as the difference between the two patent offices are somewhat stable over time.

[10] The results remain consistent when using total citations instead of only examiner-added ones (see Table A4).

The logic of the patent twin test and estimation methods

Many prior studies have assumed that the number of citations a patent receives reflects its (unobserved) quality, that is, "better" patents receive more citations. A primary goal of our patent-twins test is to isolate the effect of category contrast from any unobserved quality differences that may exist across patents. Let X_t stand for the citations the patents would receive in a year *t* in an ideal world without contrast and other non-technical influences. X_t is a property of the patented invention and, as such, is identical both at the USPTO and at the EPO. On the contrary, patent class characteristics that may affect patent citation, such as patent class contrast, size and fertility, differ between the USPTO and the EPO. We can therefore express an invention's citation counts in year t at the USPTO and its patent twin at the EPO as follows:

Eq(2): $US_{citation(t+1)} = \beta_0 + \beta_1 X_t + \beta_2 US_{classcontrast} + \beta_3 US_{classsize} + \beta_4 US_{classfertility} + \varepsilon$

Eq(3): $EPO_{citation(t+1)} = \gamma_0 + \gamma_1 X_t + \gamma_2 EPO_{classcontrast} + \gamma_3 EPO_{classsize} + \gamma_4 EPO_{classfertility} + \varepsilon$

All these quantities are observable except X_t, the unobserved quality. The goal of the patent twin design is to take X out of the picture. We achieve this by algebraically reorganizing Eq(3) to Eq(4):

$X_t = (EPO_{citation(t+1)} - \gamma_0 - \gamma_2 EPO_{classcontrast} - \gamma_3 EPO_{classsize} - \gamma_4 EPO_{classfertility})/\gamma_1 + \varepsilon$

Then we substitute this to Eq(2) and we get:

Eq(5):

$US_{citation(t+1)} = \beta_0 + \beta_1/\gamma_1(EPO_{citation(t+1)} - \gamma_0 - \gamma_2 EPO_{classcontrast} - \gamma_3 EPO_{classsize} - \gamma_3 EPO_{classfertility}) + \beta_2 US_{classcontrast} + \beta_3 US_{classsize} + \beta_4 US_{classfertility} + \varepsilon$

Note that this equation only contains observed quantities and therefore can be estimated directly. Because in the above equations the coefficient estimates are just placeholders, the equation simplifies to Eq (6) (we use α to denote that these coefficient values are not the same as in the equations above).

Eq (6):

$$US_{citation(t+1)} = \alpha_0 + \alpha_1 US_{classcontrast} + \alpha_2 US_{classsize} + \alpha_3 US_{classfertility} + \alpha_4 EPO_{citation(t+1)} - \alpha_5 EPO_{classcontrast} - \alpha_6 EPO_{classsize} - \alpha_7 EPO_{classfertility} + \varepsilon$$

We estimate equation (6), with the addition of year fixed effects and patent-dyad fixed effects to control for any additional unobserved heterogeneity. We report standard errors clustered by patent twin, but we note that the results are robust to other standard error calculations as well, such as robust standard errors (Wooldridge, 2002; Mehta et al., 2010).

Calculating the EPO variables

Twin patent's yearly patent citation count at the EPO. This variable measures the count of citations the EPO twin received at the EPO, all of which are added by examiners.

Class contrast at the EPO. The measure of patent class contrast we use in the USPTO sample cannot be directly applied to the EPO one: whereas the USPTO assigns each patent to a single primary class based on patent's "main inventive content," the EPO assigns patents to multiple primary classes. We therefore devised a close alternative measure of contrast for EPO patents. First, we calculated the recombinativeness of patents at the EPO by calculating the Herfindahl index based on their four-digit IPC class assignments. Second, for each IPC4 class in each year, we calculated the mean of the Herfindahl index of the patent applications in that year. Third, for each EPO patent, we averaged these contrast values for the classes the patent is

assigned to (for a similar approach, see Kovacs & Hannan, 2010).

Class size and class fertility at the EPO. As said, the EPO assigns patents to multiple primary classes. While the class size and class fertility variables at the USPTO are specific to each primary class, for the EPO patent twins we calculated them as the average of the size and fertility of the classes to which the patent is assigned to. As in Study 1, *Class size and class fertility* are time-varying. Table 3b reports descriptive statistics and pairwise correlations between the main variables.

---------- Insert Table 3b Here ----------

Results

Table 5 shows the results of six model specifications of Eq. (6). Models 1 and 2 are log-linear regressions using the logged count of citations in a given year as an outcome variable, while models 3 to 6 use raw citation counts as dependent variable, estimated via Poisson and negative binomial regressions, respectively. Models 1, 3, and 5 use the same specifications as presented in Study 1, while Models 2, 4, and 6 add the EPO controls to these specifications. All models include patent-twin-level and citation year fixed effects.

Before we look at the results, it may be useful to specify our priors. Based on Eq(6), we expect EPO and USPTO citation counts to be positively correlated. This is because a patent's unobserved quality should increase patent citations regardless of where a patent is classified. Additionally, we expect the other EPO class characteristics (EPO class contrast, EPO class size, EPO class fertility) to have the opposite sign as what we observe for the corresponding USPTO class characteristics. This is because, due to the transformations leading to Eq(6), those

coefficients have a minus sign. Lastly, and this is our core hypothesis, we expect the effect of USPTO patent class contrast to be positive. This is exactly what we find.

We note that, even though these models are estimated on the patent-twin subsample, the coefficient estimates in Table 5's Models 1, 3, and 5 are similar in magnitude to the estimates in Table 4's Models 2, 4, and 6. Table 5 shows that the coefficient estimate of USPTO patent class contrast on US citations is positive across all models. This effect remains statistically significant after controlling for patent class contrast at the EPO and for other EPO-level variables. This robust set of results provides compelling support for our argument that patent class contrast increases patent citations *net of any possible technical differences* across patents. Based on the negative binomial model estimates (Table 5 Model 6), one standard deviation increase in contrast leads, on average, to 3.5% more citations. Taking into consideration the full range of observed patent class contrasts (0.04 to 1.000), this translates to a 34.6% difference in predicted citations between the lowest and highest-contrast patent classes. Given that patents receive on average 15.8 citations during their lifetime of 20 years, this translates to a difference of roughly 5 citations.[11] Taken together, these findings provide compelling evidence that patent class contrast has an independent effect on patent citation over and beyond the effects that might be attributed to differences in underlying quality or other unobserved technological factors. In the Appendix (Table A2), we check the robustness of our results by estimating a range of alternative model specifications and contrast measures. The results remain consistent with those discussed above.[12]

[11] The control variables are also in line with our expectations. In line with prior findings (Carnabuci et al., 2015, Table 7), we find that the size of the primary class within which a patent is classified has a negative effect on citations. We also find a positive effect of patent class fertility on the count of citations the patent receives in a given year, in line with prior studies (Fleming, 2001; Fleming & Sorenson, 2004; Carnabuci et al., 2015).

[12] Note that the twin patents approach also allows us to test whether there is any sorting effects, e.g., whether better patents are systematically sorted into higher contrast classes. We test for sorting by estimating Eq(4) and using the predicted quality in Eq(3), which we reorganize so that US contrast is on the left-hand side. We estimated this model on a sample that contained one observation per patent. Specifically, for each patent, we took values of contrast, fertility, etc. in the granting year of the patent because that is when the sorting would happen. We did not find a

---------- Insert Table 5 Here ----------

STUDY 3

While Studies 1 and 2 demonstrate the hypothesized effect of patent class contrast on patent citations, neither study enables us to directly observe the mechanisms behind this effect. To gather more direct evidence of these mechanisms, in Study 3 we analyze an original, minute-by-minute data set on the prior art search behavior of a selected subset of patent users for which such micro-level data exists. Specifically, we analyze USPTO search log records of the queries conducted by patent examiners while searching for relevant prior art for patent applications.

Data, sample, and methods

See Figure 2 for an example of a patent examiner search log, conducted during the prior art search for patent application #11000808. The search is conducted through the EAST platform, which is available to all patent examiners and stands for Examiner Automated Search Tool. The first column "Ref" simply refers to the order of the queries conducted by the examiner within the search section. The third column "Search query" is the actual query the examiner typed in (this is akin to academics' Google Scholar searches), and the second column "Hits" shows how many results (i.e., prior patents) the system found that satisfied the search criteria. The column "DBs" refers to the databases searched; "default operator" can be "OR" or "AND;" "plurals" refers to whether the search allows for plural versions of the searched words. Finally, the last column contains the date and time when the query was conducted.

significant relationship between patent quality as predicted from the EPO sample and the category contrast of the US patent class to which the patent was classified originally. We interpret this evidence as indicating that there is no sorting effect.

---------- Insert Figure 2 Here ----------

We collected and coded the search logs of all the 610,764 queries made by UPSTO patent examiners who searched for relevant prior art for all the 17,373 patent applications during the month of February 2006. The USPTO provides the search logs in a PDF format, which then need to be scanned, cleaned, and transformed to an analyzable data format.

Our tests are at the patent application level. Note that patents are assigned to a primary class by the Office of Patent Application Processing (OPAP) before the patent is assigned to an examiner. In other words, when the examiner receives the patent application, the patent application is already classified (see the "Background information on the process of class allocation at the USPTO and the EPO" section in the Appendix for more details on the classification process at the USPTO and the EPO).

Testing the hypothesized mechanisms

We derived the central hypothesis of this paper – that patent users are more likely to cite a patent if it is classified in a high-contrast patent class – from several theoretical arguments. Below we provide a set of tests designed to probe these arguments empirically. Before commenting on the results, we first introduce the independent and control variables employed in these models and subsequently we elaborate on the specific rationale and dependent variable of each test.

Independent and control variables. The main independent variable in Table 6 is the contrast of the primary class to which the application is assigned. Contrast values are calculated in the same way as in Studies 1 and 2. As in Studies 1 and 2, we control for the *size* and *fertility* of the primary class of the patent by using the lagged values of these variables from year 2005.

We also add controls for application-level factors that may influence patent search, such as (1) the *number of queries* in the prior art search, (2) whether the patent application represents a *continuation application*, (3) has a *foreign priority*, (4) is submitted by a *small entity* (as defined by the USPTO), and (5) patent *team size*, as indicated by the count of inventors on the patent. Note that we intentionally do not control for other patent-level variables commonly used in the literature, such as scientific citations or number of claims, because these are outcomes (not antecedents) of the search process. To control for possible examiner heterogeneity, we present models with examiner random effects[13]. As all search sessions take place in February 2006, we do not need to control for year trends. Table 3c provides descriptive statistics and correlations of the variables used in the models.

---------- Insert Table 3c Here ----------

Table 6 shows the results of six tests of the proposed mechanisms behind the contrast effect. First, we theorized that patent classes are more likely to channel patent users' attention and search for prior art when category contrast is high than when it is low. One way to probe this argument is by analyzing how examiners' search strategies change as a function of patent class contrast. We expect that the higher is the contrast of a patent class, the more likely is the

[13] Note that the data sampling approach followed in this paper is not suitable to including examiner fixed effects because within the month of February 2006 (a) examiners typically work within one main class (71% of the examiners in our sample review applications only within one primary class during the observation window, 20% reviews in two primary classes, 5.5% in three, and only 2% in four or more) and (b) class contrast is constant. Yet, because the Hausman models indicate that the independence assumptions of the random effects models do not hold, in additional analyses (see Table A5) we re-estimated the models of Table 6 on the subset of patent applications that are handled by examiners who worked in more than one primary class and handled at least 10 applications. In these models, we could add examiner fixed effects. While the number of observations drops significantly, the results are mostly robust to these specifications, and the effect of contrast turns insignificant only for the last model where the DV is examiner-added citation counts. Note however that in the much larger sample in Study 2, the effect of contrast on examiner-added citation count is consistently positive and significant.

examiner to use classification-based search as opposed to searching for prior art without the aid of patent classifications (such as, for instance, by keywords). To test this argument, we built a variable labelled *Classification-based search.* This is a binary 0/1 variable that captures if examiners use classification-based searches in any of the search queries within a particular search session. For this, we focus on the "Search query" column and code whether the query is a classification-based search (e.g., "235/375.ccls") or not. "ccls" stands for "current classification," and the first query in Figure 2 means that the examiner is asking the search platform to list all prior art (patents and patent applications) that are classified in the primary class 235 and subclass 375 and whose text contain the word "wine." Model 1 in Table 6 reports the estimates of a logit model that predicts the likelihood that an examiner uses classification-based search as opposed to searching for prior art without the aid of patent classifications (see, e.g., the fourth line of the query session in Figure 2). In support of our theorized mechanism, we find that the likelihood that an examiner relies on a class-based search to find relevant prior art is higher for applications in high-contrast classes than for applications in low-contrast classes.

Second, we argued that patent users are more likely to start their prior art search by focusing on high-contrast patent classes rather than on low-contrast ones. If this argument is true, we should find that examiners are more likely to initiate their queries by searching for prior art categorized within the class itself (as opposed to prior art categorized in other classes) when a patent is classified in a high-contrast class. To test this argument, we built a variable labelled *First searched class is same as patent's primary class*. The variable is a binary 0/1 measure indicating whether, upon receiving a patent application, examiners start by searching for prior art within the patent's primary class itself. For example, for the patent application shown on Figure 2, this variable takes on the value 1 because the patent is classified in class 235 and the first

classification-based prior art search query executed by the patent examiner is within class 235 itself. Focusing on the subsample of search sessions with at least one classification-based query, Model 2 investigates whether patent examiners are more likely to start prior art searches by focusing on high-contrast classes rather than low-contrast ones. In line with our proposed argument, we find that when the primary class is high contrast, examiners are more likely to start their search within the primary class of the patent application. Conversely, at low patent class contrast, examiners are more likely to start their prior art search by focusing on other patent classes.

Third, we argued that high-contrast classes enable patent users to focus a greater part of their prior art searches within the class itself, whereas low-contrast classes require spreading attention across several classes. To examine if this argument is true, we shift the focus of analysis from where examiners start their search to where they focus their attention in the subsequent search queries. Specifically, we built a variable labelled *Proportion of classes searched that are the same as the patent's primary class* to measure the extent to which an examiner focuses her search within the boundaries of the class or, conversely, broadens the scope of her search across patent classes. This variable, like the previous one, is only defined on the subset of search sessions that contain at least one classification-based searches. Model 3 in Table 6 finds that the higher is patent class contrast, the more likely are examiners to restrict the scope of their prior art searches to patents classified within the boundaries of the patent class itself. On the contrary, the lower patent class contrast is, the more likely examiners search for relevant prior art beyond the focal class. This finding is consistent with our arguments.

Fourth, we argued that it is cognitively more efficient to search for relevant prior art in high-contrast than in low-contrast classes. We broke down this argument in three parts and tested

each. We first measured the *Amount of search*[14] conducted by examiners as the (logged) count of search queries they execute in a particular search session. If our efficiency argument is true, we should find that examiners need fewer search queries to find relevant information when examining patents categorized in high- (rather than low-) contrast classes. The estimates reported in Model 4 in Table 6 align with this expectation. Second, we built a variable labelled *Search precision*, measured as the (logged) average number of "hits" generated by a search query. We expect that when examiners search for prior art in a high-contrast class, their search queries should be more "on target" and, hence, they should return fewer "hits". For example, the search query shown in Figure 2 shows that the query in line 1 "235/375.ccls and wine" on February 2nd, 2006 returned 12 hits, while the more specific search query in line 2, which adds the condition that the word "wine" needs to be followed by "making" or "manufacture", only returns 2 hits. In line with our argument, Model 5 in Table 6 shows that search queries are on average more "on target" – they return a smaller set of hits – when examiners look for prior art in high-contrast class. In synthesis, these two last tests show that, when examiners search for prior art in a high-contrast class, they tend to make fewer search queries and each of those queries yields a narrower set of hits. While both these findings are indicative of cognitive efficiency, a more conclusive test requires checking whether these search queries also yield more prior art citations. Model 6 in Table 6 tests whether examiners searching for prior art in high contrast-classes tend to add more citations (this is the same DV as used in Study 2). In line with our argument, we find that this is the case – examiners add *more* prior art citations when searching high-contrast classes than when searching low-contrast ones.

[14] We thank an anonymous reviewer for suggesting this and the following test.

In synthesis, these tests indicate that when searching high-contrast classes, examiners search less, search more on target, and add more prior art citations per search session. This pattern of results supports the argument that search for relevant prior art is cognitively more efficient in a high-contrast than in a low-contrast class.

---------- Insert Table 6 here ----------

Table A3 in the Appendix examines the robustness of these results to alternative contrast measures; the results obtained from these specifications are similar to those discussed so far. Taken together, this set of findings provides substantial support for our theoretical arguments concerning the mechanism the effect of contrast on the amount of citations a patent receives.

DISCUSSION

Drawing insights from theories of attention and information search, this paper argued that patents are more likely to receive attention, and hence citations, if they are classified in high-contrast patent classes than if they are classified in low-contrast ones. To test this hypothesis, we carried out a series of three related studies. First, we estimated a set of panel count models with patent and year fixed effects using the entire record of utility patents granted by the USPTO between 1975 and 2010, enabling us to establish the plausibility of our hypothesis in a large-scale sample. We found a strong and robust effect of patent class contrast on patent citations. Second, to more compellingly isolate the effect of class contrast from possible confounding effects, we estimated a case-control "twin patents" model using cases where the very same invention was patented in both the USPTO and in the European Patent Office. We found that

patent class contrast continues to exhibit a sizeable and robust positive effect on patent citations even when matching patent twins in a case-control design. Third, we probed the micro-level mechanisms through which patent class contrast influences the patent citation process by analyzing minute-by-minute prior art search logs from a sample of USPTO patent examiners. In accordance with our proposed theoretical mechanism, we found that examiners are more likely to rely on high-contrast classes than on low-contrast ones when searching for relevant prior art. Furthermore, when dealing with high-contrast classes, their prior art searches tend to remain confined within the boundaries of the class itself and are more efficient, i.e., they yield more citations in less time. Conversely, when relying on low-contrast classes, patent examiners are more likely to spread their attention across other classes and less likely to find relevant prior art. Taken together, these three studies provide compelling evidence in support of our hypothesis that patents categorized in high-contrast classes tend to accrue significantly more citations that otherwise-equivalent patents categorized in low-contrast classes.

Contributions to the innovation and strategy literatures

This paper contributes to the innovation and strategy literature by extending current understanding of attentional and search dynamics in the innovation process. Prior studies examined in considerable depth the role of attention and search dynamics in the creation of inventions. Complementing this perspective, we argued that the impact of an invention does not only depend on its inherent technical quality but also on whether potential users become aware of its existence and perceive it to be useful as they search for knowledge inputs (see, e.g., Mokyr 2002; Furman & Stern 2010). In line with this view, we shifted the focus of analysis from the creators to the users of inventions and examined how processes of attention allocation and search

affect the impact of inventions *after* their creation. The results of our study reveal patent class contrast as a novel and unexplored driver of patent impact. In addition to providing a more complete picture of the drivers of patent citations, this result highlights an interesting theoretical tension with respect to the costs and benefits of boundary spanning. Whereas a key finding of earlier studies is that searching for knowledge inputs across technological boundaries leads to creating better inventions, we demonstrate that patent users are more likely to identify an invention as relevant if it is classified within the boundaries of a well-defined, high-contrast technological category. We see the contrasting effects of boundary spanning during and after the creation of inventions as a potentially generative area of inquiry and hope that our study will inspire research in this direction.

Given the importance of patents as a competitive asset in contemporary knowledge-based organizations, the finding that many technically valuable patents end up "gathering dust in the corporate legal office" (Rivette & Klein, 2000, p. 162) while others obtain a great deal of attention and shape the evolution of entire industries, has attracted a great deal of research among strategy scholars (Wang et al., 2016; Corsino et al. 2019). As Podolny and Stuart (1995, p. 1225) explain, "it is frequently observed that the "best" technologies…are not necessarily the most successful ones, and this means that technical specifications alone may not be sufficient to gauge the likelihood of technological success (Katz and Shapiro 1984; Farrell and Saloner 1985; Arthur 1988)." Taking this observation as our point of departure, our study adds new insight into the growing body of literature that examines what drives a patent's future use, and hence value as a knowledge asset, over and beyond a patent's inherent technical quality (Bikard, 2018; Bikard & Marx, 2020; Ferguson & Carnabuci, 2017; Polidoro, 2020). Integrating ideas from Zuckerman's seminal model (1999) and recent categorization research (Hannan et al., 2019), we contributed to

this line of work by showing that the number of citations accruing to a patent depends on the contrast of the patent class within which it is categorized. The arguments advanced in this paper contribute to the call for exploring the cognitive and institutional reality within which the patent citation process unfolds (Murray & O'Mahony, 2007; Furman & Stern, 2010). Observed citation patterns reflect the expert work of a large community of patent users – inventors, patent examiners, and patent attorneys. Not unlike academics, patent users face constant information overload and rely on heuristic attention allocation processes and informational cues to cope with it (Bikard, 2018). In line with this perspective, we showed that seemingly neutral and inconsequential classification decisions systematically shape where patent users focus their search for prior art and, as a result, have a systematic and sizeable impact on how many citations a patent will receive.

While we have focused on citations counts as our core outcome variable, our theoretical model may offer new insights into other important aspects of the patenting process, too. In particular, patent class contrast may affect patent acceptance time because, as we have argued, search in high-contrast categories is cognitively more efficient. Prior innovation research argued that patent acceptance time is an important dimension of the patenting process that has economic and strategic consequences (Harhoff & Wagner, 2009). While conducting a compelling test of our conjecture is beyond the scope of this paper, initial unreported analyses based on our data show that patents classified in high contrast classes are accepted faster (one standard deviation in class contrast translates into a roughly 100-120 days faster acceptance). Such a shortened processing time has downstream consequences for firm funding and survival: for example, Farre-Mensa and colleagues (2020) show that startups that get their patent accepted faster have higher employment growth and higher sales growth (see also Polidoro, 2020). In light of our

encouraging initial results, we look forward to future research that extends our model to this and other aspects of the patenting process.

The results of our analyses also have practical implications. Since economists have tended to treat patent citation counts as merely a proxy of a patent's underlying technical quality (for a review see, e.g., Jaffee & de Rassenfosse, 2017), one may wonder if the citations received by a patent because of classification decisions, as opposed to its inherent technical merits, hold any *economic* value for the patent holder. Recent research suggests that the answer to this question is affirmative. Extant studies found that the prior art citations accruing to a patent effectively *add to* a patent's economic and strategic value in at least two ways that have little to do with the patent's underlying technology: first, they amplify its appeal and perceived worth in the market for technology (Mazloumian et al., 2011) and, second, they reduce the risk of costly litigation by helping to delimit the legal boundaries that protect an invention from undue appropriation (Shane, 2008; Lampe, 2012). The evidence we presented indicates that patents are less likely to escape patent users' attention if they are classified in high-contrast patent classes than if they are classified in low-contrast ones. Thus, although a quantification of the monetary value of contrast-driven citations goes beyond the scope of our paper, it seems likely that inventions classified in high-contrast classes will on average end up generating more economic value relative to otherwise-equivalent inventions classified in low-contrast classes.

This consideration suggests a thus far unknown arbitrage opportunity for patent owners. In cases where the classification of an invention is not straightforward, patent owners have an incentive to strategically nudge it towards high-contrast patent classes and away from low-contrast ones[15]. The extent to which such strategic behaviors might be feasible in practice

[15] This prompts the question whether patent applicants can reasonably predict the future contrast of a class. We suggest that applicants could use contrast at the time of patent application to proxy contrast in the near/medium

requires further investigation. Although patent offices have rules regulating the patenting process, our field interviews with a patent attorney and several patent examiners indicated that patent attorneys and experienced patent applicants do have significant leeway in influencing where their patent will be classified. For example, they may explicitly indicate in the patent application document which "prior art unit" they would like to examine the application, or they may choose a particular framing and keyword set that emphasizes linkages with a particular patent class. By so doing, patent applicants might effectively increase the likelihood that their patents will accrue more citations and will be duly referenced as prior art by future patents.[16]

Switching to a policy perspective, it is worth noticing that a key institutional mandate of patent offices is to ensure that relevant prior art be easily found and properly acknowledged (see, e.g., USPTO's Manual of Patent Examination, 2015). While to the best of our knowledge there exists no precise estimate of how often patents fail to be duly recognized as relevant prior art, most commentators agree that the occurrence of false negatives or Type II errors in prior art referencing is likely to be significant (Marx & Fuegi, 2019). Our arguments and evidence suggest that, to reduce such errors, it is important to keep patent class contrast as high as possible across the whole classification system. Some degree of fluctuation in patent class contrast is inevitable given that technologies evolve continuously and no classification scheme can capture such evolutions perfectly and timely. However, we suggest that time-varying measures of patent

future because in the short term the contrast values are relatively stable and the 5 year autocorrelation in contrast is around 0.6. This suggests that applicants could reasonably predict contrast in the short term. This short term stability is important because most technologies typically exert most of their impact in the shorter term, e.g., the peak of the future citations is in 1-3 years after granting (Mehta et al, 2010).

[16] This argument assumes that patent assignees want to get their patents known and utilized. There could be reasons that patent owners would want to hide their patents so that others do not discover it and build on it. For example, while the exact patent technology cannot be reused without paying licensing fees (that is the main reason for getting a patent), often other firms can get ideas for their own research and patent it in a way that does not violate existing patents. Ultimately, if an inventor really wants to hide their invention hidden, they will not even patent it, such as in the famous case of the Coca Cola recipe (Lobel, 2013).

class contrast such as the ones used in this or prior studies can easily be automated and updated in a timely fashion. Patent offices, such as the USPTO and EPO, might consider using such measures and instructing patent officers to assess how their patent classification decisions may affect levels of category contrast among affected patent classes.

Contributions to the broader management literature

While we applied our theoretical arguments to the particular context of innovations, category contrast is likely to affect search dynamics in a variety of other contexts as well. The kernel of our explanatory model is quite general: we proposed that category contrast influences decision maker's focus and cognitive efficiency when searching for information. Since the scope conditions of this argument – information overload, limited attention, and bounded rationality – apply rather widely, it is easy to envision opportunities for useful theoretical extensions.

A line of future research that we consider especially promising pertains to the strategic use of categories (e.g., Pontikes 2018), particularly for platform companies for which product classifications work as an interface between firms and users. For example, Netflix classifies movies according to genres such as "Action" or "Documentaries", while Amazon.com products are organized into categories such as "Books" or "Toys." Our paper suggests that higher contrast categories are more useful for users to navigate the product space available on such platforms, which in its turn may affect user reactions as well as product positioning strategies. While other ways exist to search the product space (e.g., keyword-based search or AI-based recommendation systems), our paper is a reminder that people are boundedly rational and heavily rely on category-based heuristics to process information.

Our proposed explanatory framework also offers interesting connections to other disciplines. While popularized in the management field by Zuckerman (1999), the idea that categories delimit users' "consideration set" originates from a marketing model designed to explain how consumers choose among products (Shocker et al, 1991). Similar types of models are becoming popular among economists studying revealed preferences and price setting behaviors (see e.g., Manzini & Mariotti, 2014). By illuminating the role of categories in shaping the attention and information search of users, our paper has the potential to yield productive cross-fertilizations with these literatures. We hope that the model presented here will serve as a basis and an invitation to do more research in this direction.

REFERENCES

Alcácer J, Gittelman M (2006) Patent citations as a measure of knowledge flows: The influence of examiner citations. *The Review of Economics and Statistics*, 88(4): 774-779.

Arthur WB (1988) Competing Technologies: An Overview. In *Technical Change and Economic Theory*, edited by Giovanni Dosi, Christopher Freeman, Richard Nelson, Gerald Silverberg, and Luc Soete. London: Pinter.

Arthur WB (1989) Competing technologies, increasing returns, and lock-in by historical events. *The Economic Journal,* 99(394): 116-131.

Bikard M (2018) Made in Academia: The Effect of Institutional Origin on Inventors' Attention to Science. *Organization Science,* 29 (5): 818–36.

Bikard M, Marx M (2020). Bridging Academia and Industry: How Geographic Hubs Connect University Science and Corporate Technology. *Management Science* 66(8): 3295-3798.

Bikard M (2020) Idea twins: Simultaneous discoveries as a research tool. *Strategic Management Journal* 41(8): 1528-1543.

Bowker GC, Starr SL (2000) *Sorting Things Out: Classification and its Consequences*. MIT Press.

Campitelli G, Gobet F (2010) Herbert Simon's decision-making approach: Investigation of cognitive processes in experts. *Review of General Psychology*, *14*(4), 354-364.

Carnabuci G (2013) The distribution of technological progress. *Empirical Economics,* 44(3):1143-1154.

Carnabuci G, Bruggeman J (2009) Knowledge specialization, knowledge brokerage and the uneven growth of technology domains. *Social Forces,* 88(2): 607-642.

Carnabuci G, Operti E, Kovács B (2015) The categorical imperative and structural reproduction: dynamics of technological entry in the semiconductor industry. *Organization Science,* 26(6): 1734-1751.

Castiello U, Umiltà C (1990) Size of the attentional focus and efficiency of processing. *Acta Psychologica* 73(3):195-209.

Corsino M, M Mariani, S Torrisi (2019) Firm Strategic Behavior and the Measurement of Knowledge Flows with Patent Citations. *Strategic Management Journal,* 40(7): 1040-1069.

Cowan N (2016) *Working memory capacity*. Routledge, New York.

Dahlander, L., O'Mahony, S., & Gann, D. M. (2016). One foot in, one foot out: how does individuals' external search breadth affect innovation outcomes?. *Strategic Management Journal*, 37(2), 280-302.

Dernis H, Khan M (2004) *Triadic Patent Families Methodology*. OECD.

Eggers JP, Kaplan S (2013) Cognition and Capabilities: A Multi-Level Perspective. *Academy of Management Annals*, 7(1), 293-338.

Farre-Mensa JO, Hegde D, Ljungqvist A (2020) What is a patent worth? Evidence from the US patent "lottery". *The Journal of Finance*. 75(2) 639-682.

Farrell J, Saloner G. (1985) Standardization, compatibility, and innovation. *The RAND Journal of Economics* 16(1):70-83.

Ferguson JP, Carnabuci G (2017) Knowledge recombination and gatekeeping institutions: Does spanning knowledge boundaries really lead to more impactful innovations? *Organization Science,* 28(1): 133-151.

Fleming L (2001) Recombinant uncertainty in technological search. *Management Science,* 47(1): 117-132.

Fleming L, Mingo S, Chen D (2007) Collaborative brokerage, generative creativity, and creative success. *Administrative Science Quarterly* 52(3):443-75.

Fleming L, Sorenson O (2004) Science as a map in technological search. *Strategic Management Journal* 25(8-9):909-28.

Frakes MD, &Wasserman MF (2017) Is the time allocated to review patent applications inducing examiners to grant invalid patents? Evidence from microlevel application data. *Review of Economics and Statistics*. 99(3):550-63.

Furman JL, & Stern S (2011). Climbing atop the shoulders of giants: The impact of institutions on cumulative research. *American Economic Review*, 101(5), 1933-63.

Ghosh A, Martin X, Pennings JM, Wezel FC (2014) Ambition is nothing without focus: Compensating for negative transfer of experience in R&D. *Organization Science* 25(2):572-90.

Gigerenzer G, Todd PM (1999) *Simple heuristics that make us smart*. Oxford University Press, USA.

Greve HR, Seidel ML (2015) The thin red line between success and failure: Path dependence in the diffusion of innovative production technologies. *Strategic Management Journal* 36(4): 475-496.

Hall BH, Jaffe AB, Trajtenberg M (2005) Market value and patent citations. *RAND Journal of Economics,* 36: 16-38.

Hannan MT (2010) Partiality of memberships in categories and audiences. *Annual Review of Sociology,* 36:159–181.

Hannan MT, Le Mens G, Hsu G, Kovács B, Negro G, Pólos L, Pontikes E, Sharkey AJ (2019) *Concepts and categories: Foundations for sociological and cultural analysis.* Columbia University Press.

Hannan MT, Pólos L, Carroll GR (2007) *Logics of organization theory: Audiences, codes, and ecologies*. Princeton University Press.

Harhoff D, Wagner S (2009) The duration of patent examination at the European Patent Office. *Management Science* 55(12): 1969-1984.

Helfat CE, Peteraf MA (2015) Managerial cognitive capabilities and the microfoundations of dynamic capabilities. *Strategic Management Journal* 36(6):831-50.

Hsu G. (2006a) Evaluative schemas and the attention of critics in the US film industry. *Industrial and Corporate Change*. 15(3):467-96.

Hsu G (2006b) Jacks of all trades and masters of none: audiences' reactions to spanning genres in feature film production. *Administrative Science Quarterly,* 51:420–50.

Jaffe AB, Trajtenberg M (2002) *Patents, citations, and innovations: A window on the knowledge economy*. MIT Press.

Jaffe AB, De Rassenfosse G (2017) Patent citation data in social science research: Overview and best practices. In *Research Handbook on the Economics of Intellectual Property Law* Edward Elgar Publishing.

Katila, R, Ahuja G (2002) Something old, something new: A longitudinal study of search behavior and new product introduction. *Academy of Management Journal*, *45*(6): 1183-1194.

Katz ML, Shapiro C (1984) Network Externalities, Competition, and Compatibility. *American Economic Review* 75:424-40.

Kovács B (2017) Too hot to reject: The effect of weather variations on the patent examination process at the United States Patent and Trademark Office. *Research Policy*, 46(10): 1824-1835.

Kovács B, Hannan MT (2010) The consequences of category spanning depend on contrast. *Research in the Sociology of Organizations,* 35: 171-201.

Kuhn JM (2010) Information overload at the US Patent and Trademark Office: reframing the duty of disclosure in patent law as a search and filter problem. *Yale Journal of Law & Technology*, *13*, p.89.

Lafond F, Kim D (2019) Long-run dynamics of the US patent classification system. *Journal of Evolutionary Economics* 29:631–664.

Lampe R (2012) Strategic Citation. *Review of Economics and Statistics,* 94 (1): 320–33.

Lemley, MA (2001) Rational ignorance at the patent office. *Northwestern University Law Review*, 95(4):1-34.

Lobel O. (2013) Filing for a patent versus keeping your invention a trade secret. *Harvard Business Review*. 2013-11-21.

Manzini P, Mariotti M (2014) Stochastic Choice and Consideration Sets, *Econometrica*, 82, 1153–1176.

March JG, Simon HA (1958) *Organizations*. Wiley.

Martin X, & Mitchell W. (1998). The influence of local search and performance heuristics on new design introduction in a new product market. *Research Policy*, 26(7-8), 753-771.

Mazloumian A, Eom YH, Helbing D, Lozano S, Fortunato S. (2011) How citation boosts promote scientific paradigm shifts and Nobel prizes. *PloS One* 6(5).

Mehta A, Rysman M, Simcoe T (2010) Identifying the age profile of patent citations: new estimates of knowledge diffusion. *Journal of Applied Econometrics,* 25 (7): 1073–1222.

Mokyr J. 2002. *The Gifts of Athena: Historical Origins of the Knowledge Economy*. Princeton, NJ: Princeton University Press.

Moser P, Ohmstedt J, Rhode PW. (2018) Patent citations—An analysis of quality differences and citing practices in hybrid corn. *Management Science*. 64(4):1926-40.

Murphy, G (2004). *The big book of concepts.* MIT press.

Murray F, S O'Mahony (2007) Exploring the Foundations of Cumulative Innovation: Implications for Organization Science. *Organization Science,* 18 (6): 1006–1021.

Negro G, Hannan MT, & Fassiotto M (2015) Category signaling and reputation. *Organization Science* 26(2), 584-600.

Nelson RR, Winter SG (1982) *An Evolutionary Theory of Economic Change*. Cambridge, Ma: Belknap Press.

Newell A, Simon, HA (1972) *Human problem solving.* Englewood Cliffs, NJ: Prentice-hall.

Podolny JM, Stuart TE (1995) A Role-Based Ecology of Technological Change. *American Journal of Sociology,* 100 (5): 1224–60.

Polidoro F (2020) Knowledge, Routines, and Cognitive Effects in Nonmarket Selection Environments: An Examination of the Regulatory Review of Innovations. *Strategic Management Journal*, forthcoming.

Pontikes EG. Category strategy for firm advantage. (2018) *Strategy Science*. 3(4):620-31.

Posner MI, Petersen SE (1990) The attention system of the human brain. *Annual Review of Neuroscience*, 13(1): 25-42.

Rivette KG, Kline D. (2000) *Rembrandts in the attic: Unlocking the hidden value of patents*. Harvard Business Press.

Rosch E (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3), 192.

Rosenkopf L, Almeida P (2003) Overcoming local search through alliances and mobility. *Management Science*, 49(6): 751-766.

Rosenkopf L, & Nerkar A (2001). Beyond local search: boundary-spanning, exploration, and impact in the optical disk industry. *Strategic Management Journal*, 22(4), 287-306.

Schumpeter JA (1934) *The theory of economic development*. Cambridge, Mass.: Harvard University Press.

Shocker AD, Ben-Akiva M, Boccara B, Nedungadi P (1991) Consideration set influences on consumer decision-making and choice: Issues, models, and suggestions. *Marketing Letters*. 2(3):181-97.

Shane S (2008) *The handbook of technology and innovation management*. John Wiley & Sons.

Simon HA (1971) Designing organizations for an information-rich world. In: Martin Greenberger (ed.), *Computers, communications, and the public interest*, 37-53. Baltimore. MD: Johns Hopkins Press.

Simon HA (1986) The role of attention in cognition (pp. 105-115) in *The Brain, Cognition, and Education*. Edited by Sarah L. Friedman, Kenneth A. Klivington, Rita W. Peterson New York: Academic Press.

Stuart TE, Podolny JM (1996) Local search and the evolution of technological capabilities. *Strategic Management Journal*, 17: 21-38.

U.S. Patent and Trademark Office (2015) Manual of Patent Examination (U.S. Patent and Trademark Office, Washington, DC), https://www.uspto.gov/patent/laws-and-regulations/manual-patent-examining-procedure.

Wang, H, Zhao S, He J (2016) Increase in takeover protection and firm knowledge accumulation strategy. *Strategic Management Journal* 37(12): 2393-2412.

Wooldridge JM (2002) *Econometric Analysis of Cross Section and Panel Data*. MIT Press: Cambridge, MA.

Wry T, Lounsbury M (2013) Contextualizing the categorical imperative: Category linkages, technology focus, and resource acquisition in nanotechnology entrepreneurship. *Journal of Business Venturing,* 28: 117–133.

Wry T, Lounsbury M, Jennings PD (2014) Hybrid vigor: Securing venture capital by spanning categories in nanotechnology. *Academy of Management Journal,* 57: 1309-1333.

Zuckerman EW (1999). The categorical imperative: Securities analysts and the illegitimacy discount. *American Journal of Sociology*, 104(5): 1398-1438.

FIGURES AND TABLES

Figure 1: The evolution of category contrast for three randomly selected patent classes

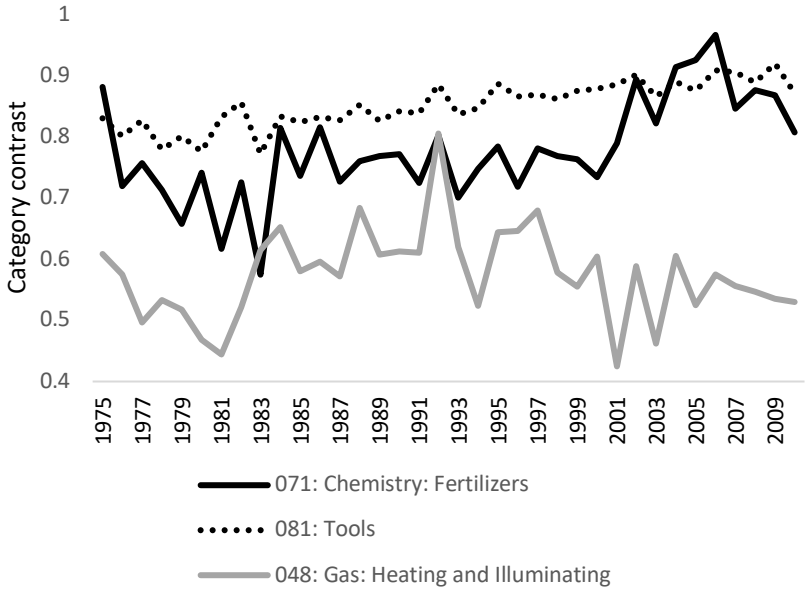

Figure 2: An example of a USPTO search log (application 11000808, search date 2006-02-06)

Ref #	Hits	Search Query	DBs	Default Operator	Plurals	Time Stamp
L1	12	235/375.ccls. and wine	US-PGPUB; USPAT; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2006/02/06 08:54
L2	2	"235"/$.ccls. and (wine adj (making or manufactur$4))	US-PGPUB; USPAT; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2006/02/06 09:00
L4	131	(extend$4 adj (handheld or hand-held or (hand adj held)))	US-PGPUB; USPAT; EPO; JPO; DERWENT; IBM_TDB	OR	OFF	2006/02/06 09:00

Table 1. Overview of the three studies

	Study 1	Study 2	Study 3
Goal of the study	To provide evidence of the effect of category contrast on future citations on the widest possible sample	To better identify the causal effect by controlling for citations to the patent twin at the EPO	To provide micro-level evidence that patent examiners use high-contrast classes to search for prior art
Observed time period	1975-2010	2001-2010	2006 February
DVs	All yearly patent citations at the USPTO	Yearly patent citations added by examiners at the USPTO	Likelihood of using patent class for prior art search Proportion of prior art search focused within class boundaries Search query length Search query precision Examiner-added citation count
Modeling approach	Log-linear, Poisson, and negative binomial models with patent and year fixed effects	Log-linear, Poisson, and negative binomial models with patent -dyad and year fixed effects	Logit and OLS models with examiner random/fixed effects
Sample	All utility patents granted between 1975-2010 that were not reclassified	All utility patents granted between 2001-2010 that were not reclassified and have a patent twin at the EPO	All prior art search conducted in 2006 February
Number of observations	N(patent)=3,260,030 N(patent-year)= 49,731,673.	N(Patent-dyad)= 67,389 N(Patent dyad-year)= 621,925.	N(search sessions)= 17,373 N(search queries)=610,764

Table 2. The primary classes with the 10 highest and lowest average contrast values between 2001-2010.

USPC primary class	Class title	Contrast	Proportion of patents in class only classified in this class	Average proportion of backward citations within class	Class size (number of patents in the primary class)
452	Butchering	0.9489	0.8965	0.7550	374
365	Static information storage and retrieval	0.9470	0.8908	0.7543	24101
439	Electrical connectors	0.9449	0.8883	0.7929	16511
84	Music	0.9435	0.8833	0.8145	3024
369	Dynamic information storage or retrieval	0.9416	0.8599	0.8398	10479
157	Wheelwright machines	0.9405	0.8857	0.6871	155
36	Boots, shoes, and leggings	0.9383	0.8807	0.7547	2224
241	Solid material comminution or disintegration	0.9349	0.8740	0.8359	2088
343	Communications: radio wave antennas	0.9334	0.8673	0.6524	7417
473	Games using tangible projectile	0.9330	0.8631	0.8425	5038
…	……	….	….	….	….
202	Distillation: apparatus	0.5666	0.2194	0.3602	217
252	Compositions	0.5561	0.2159	0.4572	5109
281	Books, strips, and leaves	0.5559	0.2836	0.4187	318
23	Chemistry: physical processes	0.5519	0.3376	0.2302	126
51	Abrasive tool making process, material, or composition	0.5194	0.2219	0.4902	711
48	Gas: heating and illuminating	0.5102	0.2045	0.2792	471
516	Colloid systems and wetting agents	0.5069	0.2232	0.2373	462
523	Synthetic resins or natural rubbers	0.4980	0.2286	0.3071	2222
95	Gas separation: processes	0.4761	0.1582	0.4169	2072
159	Concentrating evaporators	0.4255	0.0670	0.3754	197

Table 3. Descriptive statistics and correlations

(Table 3a) Study 1: Sample: All utility patents granted between 1975-2010 that were not reclassified

Variable	Mean	Std. Dev.	Min	Max	(1)	(2)	(3)
(1) Yearly citation count (USPTO) [t+1]	0.838	2.433	0.000	376.000			
(2) Primary class contrast (USPTO) [t]	0.783	0.094	0.040	1.000	0.032		
(3) Primary class size (USPTO), logged [t]	6.171	1.132	0.693	8.993	0.093	-0.024	
(4) Primary class fertility (USPTO), logged [t]	0.539	0.261	0.000	2.876	0.263	0.117	0.451

N(patent-year)=49,731,673. N(patent)=3,260,030

(Table 3b) Study 2: Sample: Patent twins that are granted at both the USPTO and EPO between 2000-2010.

Variable	Mean	Std. dev	Min	Max	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1) Yearly citation count by examiners (USPTO) [t+1]	0.384	0.815	0	24							
(2) Primary class contrast (USPTO)	0.792	0.096	0.067	1	0.053						
(3) Primary class size (USPTO), logged	6.487	1.074	0.693	8.639	0.052	-0.05					
(4) Primary class fertility (USPTO), logged	0.702	0.286	0	2.304	0.145	0.075	0.361				
(5) Yearly citation count by examiners (EPO) [t+1]	0.053	0.292	0	19	0.085	0.012	0.014	0.01			
(6) Average contrast of the patent's classes (EPO)	0.602	0.093	0.234	1	0.01	0.428	-0.066	0.166	-0.009		
(7) Average size of the patent's classes (EPO), logged	7.062	1.083	0.693	9.429	0.051	-0.184	0.432	0.358	-0.004	-0.111	
(8) Average fertility of the patent's classes (EPO), logged	0.289	0.129	0	1.946	-0.093	0.155	-0.069	-0.217	0.01	0.334	-0.276

N(dyad-year)=621,925. N(dyad)= 67,389

(Table 3c) Study 3: Sample: All prior art search conducted in 2006 February

Variable	Obs	Mean	Std. Dev.	Min	Max	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1) Primary class fertility (USPTO), logged	17,373	-0.45	0.43	-2.64	2.91								
(2) Primary class size (USPTO), logged	17,373	4.11	1.39	0.00	6.84	0.05							
(3) Primary class contrast (USPTO)	17,373	0.78	0.12	0.22	1.00	-0.06	0.17						
(4) Number of queries in the search session	17,373	3.01	1.10	0.69	6.72	0.07	-0.04	-0.07					
(5) Average number of hits per search, logged	17,373	9.64	4.22	0.00	29.35	0.04	0.02	-0.05	0.57				
(6) Classification-based search used	17,373	0.69	0.46	0.00	1.00	0.04	-0.01	0.06	0.41	0.26			
(7) Proportion of classification-based search queries to all queries	11,210[b]	0.71	0.33	0.00	1.00	0.03	0.07	0.15	-0.11	-0.05	N/A[a]		
(8) Whether the first classification-based search is within the patent's primary class	11,210[b]	0.81	0.39	0.00	1.00	0.03	0.09	0.11	-0.07	-0.03	N/A[a]	0.61	
(9) Count of examiner added citations	12,984[c]	7.14	6.45	0.00	97.00	0.05	-0.11	-0.01	0.13	0.04	0.06	-0.02	-0.03

[a] Correlation is not defined because this variable is only defined for search session with at least one classification-based search.
[b] Only defined for search sessions with at least one classification based search
[c] Only available for granted patents

Table 4: Study 1, Patent fixed effects models on yearly citation counts. See Table A1 for robustness checks.

	M1	M2	M3	M4	M5	M6
Model type	Log-linear		Poisson		Negative binomial	
$DV_{(t+1)}$	ln(citation count)		Citation count		Citation count	
$Contrast_t$	0.1658	0.0778	0.8007	0.1784	0.5053	0.2764
	(0.0016)	(0.0017)	(0.0043)	(0.0044)	(0.0048)	(0.0048)
$Ln(classsize_t)$		-0.0074		0.0592		0.0022
		(0.0002)		(0.0005)		(0.0005)
$Ln(fertility_t)$		0.4853		1.2520		0.8891
		(0.0008)		(0.0013)		(0.0017)
Constant	0.2460	0.2358			-0.0362	-0.1336
	(0.0019)	(0.0022)			(0.0074)	(0.0079)
N	48,309,345	48,309,345	48,309,345	48,309,345	48,309,345	48,309,345
Log-likelihood	-2.8458e+07	-2.8241e+07	-4.4969e+07	-4.4506e+07	-4.0952e+07	-4.0820e+07

Year range: 1975-2010. All models include patent fixed effects and citing year fixed effects. Standard errors (in parentheses) clustered on patents.

Table 5: Study 2, Patent-twin tests: yearly forward citations added by examiners at the USPTO as a function of primary class contrast at the USPTO, while controlling for citations to the same technology (patent twin) at the EPO. See Table A2 for robustness checks.

	M1	M2	M3	M4	M5	M6
Model type	Log-linear		Poisson		Negative binomial	
$DV_{(t+1)}$	US_ln(citations)		US Count of citations		US Count of citations	
$US\_Contrast_t$	0.1076	0.1028	0.2978	0.3243	0.3298	0.3686
	(0.0156)	(0.0156)	(0.0746)	(0.0746)	(0.0760)	(0.0757)
$US\_Ln(classsize_t)$	-0.0644	-0.0586	-0.0864	-0.0793	-0.0663	-0.0737
	(0.0019)	(0.0019)	(0.0079)	(0.0081)	(0.0080)	(0.0081)
$US\_Ln(fertility_t)$	0.1094	0.1043	0.5847	0.4764	0.6040	0.4828
	(0.0053)	(0.0054)	(0.0226)	(0.0233)	(0.0231)	(0.0241)
$EPO\_citationcount_{(t+1)}$		0.0354		0.0863		0.0888
		(0.0017)		(0.0055)		(0.0060)
$EPO\_classsizeyear_t$		0.0187		0.3209		0.1963
		(0.0042)		(0.0208)		(0.0154)
$EPO\_contrast_t$		-0.2985		-0.4954		-0.2938
		(0.0201)		(0.1040)		(0.1039)
$EPO\_fertility_t$		-0.1112		-0.5590		-0.7107
		(0.0105)		(0.0555)		(0.0555)
Constant	0.6146	0.6373			1.7205	0.7765
	(0.0179)	(0.0358)			(0.0889)	(0.1398)
N	621,925	621,925	536,311	536,311	536,311	536,311
Log-likelihood	-193,870	-193,402	-312,494	-312,067	-310,859	-310,502

Year range: 2001-2010. All models include patent-twins fixed effects and citing year fixed effects. Standard errors (in parentheses) clustered on patents.

Table 6. Study 3: Logit and OLS models on whether the search query contain classification-based search, how contrast influences the amount the examiner searches, the number of hits these searches return, and the count of the citations examiners end up adding. See Tables A3 and A5 for robustness checks.

	M1	M2	M3	M4	M5	M6
DV	Use class	First class same as main class	Prop class-based searches within main	Ln(total line +1)	Ln(hit count + 1)	Ln(examiner added citation count+1)
Type of model	Logit	Logit	OLS	OLS	OLS	OLS
Contrast	4.8058	3.8270	0.7776	-0.4393	-1.5320	3.8264
	(0.5432)	(0.3835)	(0.0489)	(0.1412)	(0.5035)	(0.9620)
Ln(totalline)	1.4466	-0.2177	-0.0394		0.0304	0.0163
	(0.0365)	(0.0313)	(0.0034)		(0.0006)	(0.0012)
Ln(classsize)	-0.0149	0.0925	-0.0001	-0.0149	0.0756	-0.5132
	(0.0355)	(0.0241)	(0.0030)	(0.0092)	(0.0331)	(0.0633)
Ln(fertility)	0.1141	0.1063	0.0146	0.0929	0.2458	0.7596
	(0.1084)	(0.0771)	(0.0094)	(0.0284)	(0.1014)	(0.1905)
Continuation patent	-0.2995	-0.1285	-0.0124	-0.0351	-0.1590	-0.6854
	(0.0573)	(0.0566)	(0.0061)	(0.0139)	(0.0468)	(0.0977)
Has foreign priority	0.0828	0.0689	0.0082	-0.0524	0.2196	-0.7132
	(0.0594)	(0.0583)	(0.0062)	(0.0142)	(0.0479)	(0.1010)
Small entity	-0.0574	0.0043	-0.0178	-0.0695	-0.2936	-0.2279
	(0.0869)	(0.0853)	(0.0092)	(0.0209)	(0.0702)	(0.1349)
Inventor count	-0.0374	-0.0110	0.0010	0.0032	-0.0075	-0.0211
	(0.0146)	(0.0146)	(0.0016)	(0.0036)	(0.0122)	(0.0253)
Constant	-6.3321	-0.9156	0.2232	3.5026	9.5195	6.7722
	(0.4601)	(0.3318)	(0.0421)	(0.1156)	(0.4140)	(0.7875)
N	17,366	11,204*	11,204*	17366	17364	12978

All models contain examiner random effects. Robust standard errors in parentheses.
* Note that Models 2 and 3 are run on the subset of patent applications in which the examiner conducts at least one classification-based search.

APPENDIX

Results with alternative operationalizations of the contrast variable

Moving average

The set of patents based on which we calculate patent class contrast are changing yearly, but one may argue that contrast has a longer-lasting effect and that past values of contrast may also matter for attention. Therefore, we also calculate two alternative, smoothened versions of the contrast values that take past into account. First, we calculate a smoothened version by taking a three-year moving average of the contrast of the class, weighted by the number of patents in that year and class. For example, if class 205 has a 100 patents and a contrast value of 0.8 in 2005, 50 patents and contrast of 0.7 in 2004, and 40 patents and 0.75 contrast in 2003, then the weighted moving average for year 2005 will be (0.8*100+0.7*50+0.75*40)/(100+50+40)=0.763. Note that these values will be left censored and undefined for years 1975 and 1976.

LOWESS smoothing

As a second alternative, we calculate the LOWESS smoothed values of contrast. The LOWESS smoothening (which stands for locally weighted scatterplot smoothing), is a local regression-based smoothing (see Cleveland & Devlin 1988), which takes a moving window of observations and estimates a linear regression and then instead of the observed value, for smoothing it uses the values predicted by the regression. We use the LOWESS smoothing with 0.25 bandwidth, meaning that the observations used for the smoothing regression use 25% [moving window] of the contrast values for each class.

Proportion of patents in the main class that are only classified into that class

A third alternative operationalization to capture class contrast is the "Proportion of patents in the main class that are only classified into that class."

Average proportion of backward citations to patents within the primary class

A fourth alternative way to capture class contrast is to calculate the average proportion of backward citations to patents within the same primary class. This measure, by definition, can only use patents that have at least one backward citations.

Please see the pairwise correlation values below.

	Contrast	Contrast_ movingavg3	Contrast_ lowess	Contrast_ binary	Contrast_ mean_propbwciteswithin
Contrast	1				
Contrast_ movingavg3	0.8408	1			
Contrast_ lowess	0.8524	0.9064	1		
Contrast_ binary	0.9051	0.7994	0.788	1	
Contrast_ mean_propbwciteswithin	0.3381	0.361	0.3383	0.3383	1

Table A1: Robustness checks to Study 1. Patent fixed effects models on yearly citation counts, using alternative specifications to the contrast measure.

	M1	M2	M3	M4	M5	M6
Model type	Log-linear	Poisson	NBREG	Log-linear	Poisson	NBREG
$DV_{(t+1)}$	ln(cites)	cites	cites	ln(cites)	cites	cites
Contrast measure	Moving average of contrast values in years t, t-1, and t-2			LOWESS smoothing with 0.25 bandwidth		
$Contrast_t$	0.1004	0.2424	0.3356	0.1449	0.3551	0.4314
	(0.0021)	(0.0055)	(0.0057)	(0.0032)	(0.0086)	(0.0072)
$Ln(classsize_t)$	-0.0108	0.0472	-0.0057	-0.0099	0.0439	-0.0036
	(0.0002)	(0.0005)	(0.0005)	(0.0002)	(0.0005)	(0.0005)
$Ln(fertility_t)$	0.4845	1.2573	0.8946	0.4850	1.2537	0.8955
	(0.0008)	(0.0014)	(0.0018)	(0.0008)	(0.0015)	(0.0018)
Constant	0.2309		-0.0598	0.2011		-0.1937
	(0.0023)		(0.0074)	(0.0029)		(0.0088)
N	44,784,368	44,784,368	44,784,368	44,974,047	44,974,047	44,974,047
Log-likelihood	-2.6069e+07	-4.0939e+07	-3.7684e+07	-2.6167e+07	-4.1095e+07	-3.7827e+07
	M7	M8	M9	M10	M11	M12
Model type	Log-linear	Poisson	NBREG	Log-linear	Poisson	NBREG
$DV_{(t+1)}$	ln(cites)	cites	cites	ln(cites)	cites	cites
Contrast measure	Proportion of patents only classified into this main class			Average proportion of backward citations to patents within the primary class		
$Contrast_t$	0.0311	0.0672	0.1053	0.1155	0.4418	0.3775
	(0.0009)	(0.0024)	(0.0026)	(0.0015)	(0.0044)	(0.0042)
$Ln(classsize_t)$	-0.0041	0.0665	0.0090	-0.0057	0.0625	0.0065
	(0.0002)	(0.0005)	(0.0005)	(0.0002)	(0.0005)	(0.0005)
$Ln(fertility_t)$	0.4840	1.2532	0.8877	0.4830	1.2480	0.8893
	(0.0008)	(0.0013)	(0.0017)	(0.0008)	(0.0013)	(0.0017)
Constant	0.2465		0.0401	0.1877		-0.1577
	(0.0016)		(0.0070)	(0.0018)		(0.0075)
N	48,288,569	48,288,569	48,288,569	48,288,569	48,288,569	48,288,569
Log-likelihood	-2.8235e+07	-4.4493e+07	-4.0810e+07	-2.8232e+07	-4.4488e+07	-4.0807e+07

Year range: 1975-2010. All models include patent fixed effects and citing year fixed effects. Standard errors (in parentheses) clustered on patents.

Table A2: Robustness checks to Study 2, Patent-twin tests: yearly forward citations added by examiners at the USPTO as a function of primary class contrast at the USPTO, while controlling for citations to the same technology (patent twin) at the EPO, using alternative specifications to the contrast measure.

	M1	M2	M3	M4	M5	M6
Model type	Log-linear	Poisson	NBREG	Log-linear	Poisson	NBREG
$DV_{(t+1)}$	US_ln(cites)	US cites	US cites	US_ln(cites)	US cites	US cites
Contrast measure	Moving average of contrast values in years t, t-1, and t-2			LOWESS smoothing with 0.25 bandwidth		
$US\_Contrast_t$	0.1567	0.5857	0.6194	0.2181	0.6247	0.6689
	(0.0155)	(0.0936)	(0.0910)	(0.0326)	(0.1824)	(0.1539)
$US\_Ln(classsize_t)$	-0.0513	-0.1271	-0.1043	-0.0506	-0.1124	-0.0943
	(0.0015)	(0.0083)	(0.0081)	(0.0015)	(0.0077)	(0.0076)
$US\_Ln(fertility_t)$	0.0605	0.3681	0.3804	0.0608	0.3718	0.3831
	(0.0043)	(0.0229)	(0.0233)	(0.0043)	(0.0230)	(0.0233)
$EPO\_citationcount_{(t+1)}$	0.0226	0.0711	0.0712	0.0226	0.0706	0.0708
	(0.0013)	(0.0055)	(0.0059)	(0.0013)	(0.0054)	(0.0059)
$EPO\_classsizeyear_t$	0.0000	0.0001	0.0001	0.0000	0.0001	0.0001
	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)	(0.0000)
$EPO\_contrast_t$	-0.2610	-0.4744	-0.4249	-0.2673	-0.5324	-0.4738
	(0.0146)	(0.0909)	(0.0906)	(0.0146)	(0.0906)	(0.0906)
$EPO\_fertility_t$	-0.0417	-0.4241	-0.4308	-0.0408	-0.4224	-0.4297
	(0.0032)	(0.0262)	(0.0270)	(0.0032)	(0.0262)	(0.0270)
Constant	0.5456		1.9490	0.4960		1.8712
	(0.0177)		(0.1056)	(0.0280)		(0.1387)
N	1,074,312	827,866	827,866	1,074,372	827,947	827,947
Log-likelihood	-224137.4386	-433178.5289	-431706.9644	-224210.8333	-433270.4487	-431792.7607

(CONTINUED ON NEXT PAGE)

(CONTINUED FROM PREVIOUS PAGE)

	M7	M8	M9	M10	M11	M12
Model type	Log-linear	Poisson	NBREG	Log-linear	Poisson	NBREG
$DV_{(t+1)}$	US_ln(cites)	US cites	US cites	US_ln(cites)	US cites	US cites
Contrast measure	Proportion of patents only classified into this main class			Average proportion of backward citations to patents within the primary class		
$US\_Contrast_t$	0.0555	0.2228	0.2341	0.0439	0.2542	0.3196
	(0.0082)	(0.0486)	(0.0418)	(0.0147)	(0.0948)	(0.0784)
$US\_Ln(classsize_t)$	-0.0576	-0.0779	-0.0725	-0.0582	-0.0811	-0.0751
	(0.0019)	(0.0099)	(0.0081)	(0.0019)	(0.0099)	(0.0080)
$US\_Ln(fertility_t)$	0.0999	0.4749	0.4823	0.1018	0.4817	0.4885
	(0.0056)	(0.0289)	(0.0241)	(0.0056)	(0.0289)	(0.0240)
$EPO\_citationcount_{(t+1)}$	0.0351	0.0863	0.0888	0.0351	0.0863	0.0888
	(0.0022)	(0.0067)	(0.0060)	(0.0022)	(0.0067)	(0.0060)
$EPO\_classsizeyear_t$	0.0149	0.3208	0.1965	0.0144	0.3166	0.1968
	(0.0039)	(0.0249)	(0.0154)	(0.0039)	(0.0249)	(0.0154)
$EPO\_contrast_t$	-0.2978	-0.4883	-0.2916	-0.3029	-0.5127	-0.3163
	(0.0180)	(0.1213)	(0.1039)	(0.0180)	(0.1213)	(0.1043)
$EPO\_fertility_t$	-0.1051	-0.5575	-0.7091	-0.1057	-0.5587	-0.7052
	(0.0094)	(0.0633)	(0.0555)	(0.0094)	(0.0633)	(0.0555)
Constant	0.6928		0.9283	0.7067		0.8980
	(0.0323)		(0.1271)	(0.0328)		(0.1327)
N	656,443	536,309	536,309	656,443	536,309	536,309
Log-likelihood	-186955.6357	-312062.3843	-310499.0143	-186976.6368	-312071.9667	-310506.4020

Year range: 2001-2010. All models include patent fixed effects and citing year fixed effects. Standard errors (in parentheses) clustered on patents.

Table A3. Robustness check to Study 3: Logit and OLS models on whether the search query contain classification-based search, how contrast influences the amount the examiner searches, the number of hits these searches return, and the count of the citations examiners end up adding. Compares to Table 6 but models in this table use alternative specifications to the contrast measure.

		M1	M2	M3	M4	M5	M6
		DV					
		Use class	First class same as main class	Prop class-based searches within main	Ln (total line +1)	Ln (hit count + 1)	Ln(examiner added citation count+1)
		Type of model					
		Logit	Logit	OLS	OLS	OLS	OLS
Alternative contrast measures used	Contrast calculated as class size-weighted moving average of contrast values in years t, t-1, and t-2.	5.1745 (0.5557)	4.1463 (0.3914)	0.8119 (0.0499)	-0.4104 (0.1443)	-1.5553 (0.5152)	4.1049 (0.9827)
	Contrast calculated as LOWESS smoothened contrast with alpha=0.25	5.5830 (0.5826)	4.3213 (0.4148)	0.8445 (0.0526)	-0.4551 (0.1524)	-1.7249 (0.5448)	4.2428 (1.0266)
	Contrast calculated as Proportion of patents only classified into this main class	2.8127 (0.2892)	2.0138 (0.2104)	0.4044 (0.0267)	-0.2262 (0.0762)	-0.8027 (0.2723)	2.0726 (0.5162)
	Contrast calculated as the Average proportion of backward citations to patents within the primary class	1.9731 (0.3970)	1.3570 (0.2703)	0.3615 (0.0344)	-0.7850 (0.1008)	-2.1748 (0.3618)	0.7285 (0.6884)

Notes: Each cell is based on a separate regression model and shows the effect of contrast on the respective dependent variables. All models contain examiner random effects and the set of controls used in Table 6. Estimates for the control variables are not shown here but full estimates tables are available from the authors. Robust standard errors in parentheses.

Table A4: Study 2, Patent-twin tests, DV: count of total yearly forward citations as a function of primary class contrast at the USPTO, while controlling for citations to the same technology (patent twin) at the EPO.

	M1	M2	M3	M4	M5	M6
Model type	Log-linear		Poisson		Negative binomial	
$DV_{(t+1)}$	US_ln(citations)		US Count of citations		US Count of citations	
$US\_Contrast_t$	0.1042	0.1151	0.2840	0.3305	0.2026	0.2132
	(0.0217)	(0.0217)	(0.0427)	(0.0427)	(0.0444)	(0.0454)
$US\_Ln(classsize_t)$	-0.0318	-0.0327	-0.0503	-0.0598	-0.0448	-0.0481
	(0.0026)	(0.0026)	(0.0046)	(0.0047)	(0.0045)	(0.0047)
$US\_Ln(fertility_t)$	0.4126	0.3747	0.9633	0.8468	0.6659	0.5879
	(0.0074)	(0.0076)	(0.0113)	(0.0118)	(0.0133)	(0.0141)
$EPO\_citationcount_{(t+1)}$		0.0533		0.0743		0.0955
		(0.0024)		(0.0033)		(0.0046)
$EPO\_classsizeyear_t$		0.0823		0.2273		0.0099
		(0.0059)		(0.0122)		(0.0064)
$EPO\_contrast_t$		0.0524		0.5759		0.4224
		(0.0280)		(0.0599)		(0.0589)
$EPO\_fertility_t$		-0.1037		-0.5429		-0.7085
		(0.0147)		(0.0323)		(0.0358)
Constant	0.3010	-0.2684			0.2410	0.1456
	(0.0248)	(0.0498)			(0.0480)	(0.0675)
N	621925	621925	621925	621925	621925	621925
Log-likelihood	-398989	-398417	-630714	-629833	-582008	-581561

Year range: 2001-2010. All models include patent-dyad fixed effects and citing year fixed effects. Standard errors (in parentheses) clustered on patents.

Table A5. Study 3: Logit and OLS models on whether the search query contain classification-based search, how contrast influences the amount the examiner searches, the number of hits these searches return, and the count of the citations examiners end up adding. This table compares to Table 6, but this models are with examiner FE and are estimated on the subsample of patent examiners who handled patents in at least two different primary classes and handled at least 10 applications.

	M1	M2	M3	M4	M5	M6
DV	Use class	First class same as main class	Prop class-based searches within main	Ln(total line +1)	Ln(hit count + 1)	Ln(examiner added citation count+1)
Type of model	Logit	Logit	OLS	OLS	OLS	OLS
Contrast	4.1660	3.8115	0.3363	-0.6528	-2.8090	0.4412
	(1.6688)	(1.3437)	(0.1661)	(0.3611)	(1.0263)	(2.4668)
Ln(totalline)	1.2397	-0.3238	-0.0561		2.0073	0.5935
	(0.0872)	(0.0801)	(0.0092)		(0.0554)	(0.1303)
Ln(classsize)	0.0717	-0.0545	-0.0120	0.0285	0.0644	-0.1833
	(0.1080)	(0.0845)	(0.0104)	(0.0243)	(0.0689)	(0.1695)
Ln(fertility)	0.2394	-0.3569	0.0098	-0.1337	0.3012	0.9592
	(0.3119)	(0.2395)	(0.0269)	(0.0649)	(0.1845)	(0.4388)
Continuation patent	-0.3050	-0.3083	-0.0291	0.0016	-0.1784	-0.7529
	(0.1355)	(0.1302)	(0.0152)	(0.0344)	(0.0977)	(0.2291)
Has foreign priority	0.0811	0.0628	0.0223	-0.0683	0.1135	-0.6035
	(0.1437)	(0.1322)	(0.0155)	(0.0352)	(0.1001)	(0.2373)
Small entity	-0.2816	0.2608	-0.0222	-0.0658	-0.3384	-0.3500
	(0.2176)	(0.2040)	(0.0230)	(0.0525)	(0.1492)	(0.3284)
Inventor count	0.0107	0.0341	0.0016	0.0023	-0.0360	0.0123
	(0.0309)	(0.0298)	(0.0036)	(0.0084)	(0.0238)	(0.0555)
Constant			0.6431	3.2713	6.0655	6.6717
			(0.1472)	(0.3047)	(0.8844)	(2.1212)
N	1798	1601	1889	2851	2850	2238

All models contain examiner fixed effects. Robust standard errors in parentheses.
* Note that Models 2 and 3 are run on the subset of patent applications in which the examiner conducts at least one classification-based search.

Table A6: Study 2, Patent-twin tests: total yearly forward citations at the USPTO as a function of primary class contrast at the USPTO, while controlling for citations to the same technology (patent twin) at the EPO. See Table A2 for robustness checks.

	M1	M2	M3	M4	M5	M6
Model type	Log-linear		Poisson		Negative binomial	
$DV_{(t+1)}$	US_ln(citations)		US Count of citations		US Count of citations	
$US\_Contrast_t$	0.1042	0.1151	0.2840	0.3305	0.2026	0.2132
	(0.0217)	(0.0217)	(0.0427)	(0.0427)	(0.0444)	(0.0454)
$US\_Ln(classsize_t)$	-0.0318	-0.0327	-0.0503	-0.0598	-0.0448	-0.0481
	(0.0026)	(0.0026)	(0.0046)	(0.0047)	(0.0045)	(0.0047)
$US\_Ln(fertility_t)$	0.4126	0.3747	0.9633	0.8468	0.6659	0.5879
	(0.0074)	(0.0076)	(0.0113)	(0.0118)	(0.0133)	(0.0141)
$EPO\_citationcount_{(t+1)}$		0.0533		0.0743		0.0955
		(0.0024)		(0.0033)		(0.0046)
$EPO\_classsizeyear_t$		0.0524		0.5759		0.4224
		(0.0280)		(0.0599)		(0.0589)
$EPO\_contrast_t$		0.0823		0.2273		0.0099
		(0.0059)		(0.0122)		(0.0064)
$EPO\_fertility_t$		-0.1037		-0.5429		-0.7085
		(0.0147)		(0.0323)		(0.0358)
Constant	0.3010	-0.2684			0.2410	0.1456
	(0.0248)	(0.0498)			(0.0480)	(0.0675)
N	621925	621925	621925	621925	621925	621925
Log-likelihood	-398989	-398417	-630714	-629833	-582008	-581561

Year range: 2001-2010. All models include patent-twin fixed effects and citing year fixed effects. Standard errors (in parentheses) clustered on patents.

Background information on the process of class allocation at the USPTO and the EPO

The process is essentially the same at the USPTO and at the EPO: after the application is received by the patent office, a pre-sorting office allocates the patent application to an examination unit (called "art unit" at the USPTO and "examination division" at the EPO) based on the content of the patent. This office also sets a preliminary classification for the application. The pre-sorting and preliminary classification has been typically done by trained classification experts (who often had been examiners themselves) who since around 2000 are aided by automatic text classification programs. In the very recent years, more and more of the pre-sorting and pre-classification tasks are done by Artificial Intelligence programs: the program makes a recommended classification, which the "human" examiners approve. Once the application is assigned to the examination unit, the unit can either take it on (which happens in most of the cases) or they have an option to say that they are not the expert in the subject matter and send it back for presorting (with recommendation for alternative examination units). If the art unit sends back the application, the pre-sorting office takes a more careful view at the application and sends it to an examination unit again. When the application is taken on by the examination unit, the head of the unit assigns the application to an examiner, who then does the patent examination, the search for prior art, and makes a decision about patentability. If and when the patent is approved for granting, the examiner is asked to take a careful look at the classification again and may review the original classification provided by the pre-sorting office – slight or major changes to the classification may be needed partly because the original classification is done cursorily so may not be perfect, but also because the content of the patent may have changed during the examination process for example as the claim for novelty was narrowed by the examiner. Approving the final classification is an important part of the examiners' task. In both the USPTO and the EPO offices, there is a quality assurance office which randomly selects a few percentages (3-4%) of the approved classification and reviews whether the examiner has done a good job.