Does Copyright Affect Reuse? Evidence from the Google Books Digitization Project

Abhishek Nagaraj [*]
nagaraj@mit.edu

September 11, 2014

Abstract

While digitization projects like Google Books have dramatically increased access to information, this study examines how the ability to reuse information and diffuse knowledge to a wider audience depends crucially on features of copyright law. I use the digitization of both copyrighted and non-copyrighted issues of *Baseball Digest*, a publication digitized under Google Books to measure the impact of copyright on a prominent venue for reuse: Wikipedia. A specific feature of the 1909 Copyright Act, *copyright renewal requirements* ensure that material published before 1964 is out of copyright, which allows causal estimation of the impact of copyright on Wikipedia across this sharp cutoff. Estimates suggest that the Google Books digitization event caused a significant increase in information on Wikipedia pages, but copyright hurt the extent of diffusion as measured by citations as well as the reuse of images and text. The negative impact of copyright on diffusion is more pronounced for images than for text, and for topics that have few alternate sources of information. Information deficiencies due to copyright have real impacts on readership – affected pages have on average 30-80% less internet traffic than pages unaffected by copyright.

[*] MIT Sloan School of Management. E62-368, 100 Main Street, Cambridge, MA 02139. I am thankful to Erik Brynjolfsson, Daniel Fehder, Jeff Furman, Shane Greenstein, Joshua Krieger, Matt Marx, Aruna Ranganathan, Fabian Waldinger, Pai-Ling Yin, participants of the Economic Sociology Working Group and MIT Economics Third Year Lunch and especially Pierre Azoulay, Scott Stern, Catherine Tucker and Heidi Williams for comments. All errors remain mine.

1 Introduction

Copyright law regulates the production, use, and reuse of creative work, and yet there is little empirical evidence available to guide the design of copyright policy. A copyright report by the Republican Study Committee (Khanna, 2012) states the basic problem quite succinctly:

> We frankly may have no idea how [copyright] actually hurts innovation, because we don't know what isn't able to be produced as a result of our current system.

This paper proposes a framework for estimating the impact of copyright on creative reuse in the light of a counterfactual. What impact copyright has on the diffusion of ideas is a central question in the "core copyright industries" which include newspapers, books, cinema, music, radio, television and software. In the United States alone, these industries employ over 5.4 million employees, and their GDP share is estimated to be at 6.6% (Siwek, 2006) and growing at a rate of 46% annually (US Department of Commerce, 2012).

Previous work in copyright has been largely theoretical and has focused on estimating the elasticity of new creative products with the strength of copyright protection (Varian (2006), Watt and Towse (2006)). However, because the creative process is cumulative in many markets (Williams (2013), Furman and Stern (2011), Scotchmer (1991)), copyright may also affect reuse by changing the cost of reusing creative inputs in the production of new knowledge (Li et al. (2012), Lerner et al. (2012)). This second impact is the focus of this study.

According to existing theory, whether copyright impedes or promotes the reuse of knowledge is unclear. Prospect theory (Kitch, 1977; Hardin, 1968; Landes and Posner, 2002; Liebowitz and Margolis, 2004) holds that broad and strong intellectual property rights are useful, because they both incentivize the creation of new knowledge (Gans, 2014) and spur reuse by reducing contracting costs for the licensing of existing knowledge by stimulating the maintenance of works already created (Mazzoleni and Nelson (1998), Gallini and Scotchmer (2002), Gans and Stern (2010), Arora et al. (2001)). However, under a competing logic, copyright imposes transaction costs that prevent the exchange of information. These transaction costs are likely to be significant given the inherent uncertainty in the value of information (Arrow, 1962) and the increasingly digital nature of knowledge production including declining costs of access and reuse (Lessig, 2005; Benkler, 2006; Zittrain, 2009; Lemley, 2004). While diametrically opposed, these two theories are clear in their respective predictions; according to one theory, copyright lubricates the market for ideas, and, according to the other, it impedes the free diffusion of ideas.

Given contradictory theoretical predictions, systematic empirical analysis of the impact of copyright on reuse could help make progress on this question. Unfortunately, such analysis has largely been absent in the innovation and management literature on the topic, perhaps due to empirical challenges. Issues of data and measurement are central. Unlike scientific output which can be measured through patents and the publication record, the diffusion of cultural information is more informal and harder to track. When data can be located, copyright applies to all information by default and persists often for over a hundred years. This, makes it rather difficult to observe diffusion in the absence of copyright. Finally, even if these problems can be resolved, comparing reuse of copyrighted and non-copyrighted information is challenging because unobserved variables like the underlying quality of information are often correlated with reuse. Put another way, we do not know if information was reused because it was off-copyright or because it was inherently of higher quality.

In this paper, I exploit a natural experiment that occurred during a marquee digitization project in the history of the internet: the digitization of about 30 million works by Google Books. During the digitization project in December 2008, Google Books digitized all existing issues (from 1940 to 2008) of *Baseball Digest*, a prominent baseball magazine, and made them available online to readers for free. Apart from the fact that it is perhaps one of the most important reference sources on the game of baseball, *Baseball Digest* is particularly useful for my purposes because issues published before 1964 are in the public domain. This is the case, due to an accidental failure to renew copyrights, which was a poorly understood legal requirement for works published in this period. Consequently, pre-1964 *Baseball Digest* issues can be legally reused, while those published after 1964 are copyrighted and cannot be legally reused (but they can be read on Google Books). This study exploits this idiosyncratic variation within the copyright status of a single publication to measure the impact of copyright on reuse. In particular, I focus on the reuse of magazine material on Wikipedia, a natural venue to investigate this question. Not only is Wikipedia the fifth most visited website on the internet (receiving about 10 billion page-views every month) as well as a natural venue to read about the history of baseball, it also stores all past versions of a given page, allowing the analyst to track how information changes in response to the Google Books digitization event.

To make things clearer, consider the example in Figure 1. The example shows scanned pages from Google Books of two issues of *Baseball Digest.* One of them (Panel A) is about a feature on Felipe Alou published in 1963, rendering it out of copyright. The other (Panel B) is on Johnny Callison published in 1964, making it in copyright. For these two players, Figure 1 depicts pages on Wikipedia as they appeared in December 2013. Neither of these two pages displayed the players' images before *Baseball Digest* was digitized, but in 2013, while Alou's page had an image from Baseball Digest, Callison's page had no images at all. Despite being printed in two issues that were published around the same time, one image finds use in a broader context while the other seems lost among the pages of Baseball Digest. Further analysis shows that the page for Alou contains a citation to *Baseball Digest*, has a slightly greater amount of text, and has experienced about a 121% increase in traffic since 2008 as compared to only a 23% increase for Callison's page. While many alternative explanations (such as differences in player popularity) could account for the differences, the statistical analysis isolates the role of copyright in establishing these patterns.

Specifically, the estimation proceeds as follows. First, I identify a set of about five hundred prominent baseball players and a comparable set of basketball players (to act as an additional layer of control) who were active between 1944 and 1984. For this set of over 1000 players, I collect data from 13 different versions of their Wikipedia pages, one each from years 2001 to 2013. For each of these page-year observations, I count the number of images, the words of text, and citations to *Baseball Digest.* I also collect data on internet traffic, which is available from 2007 to 2013. All players who made their debut before 1964 form the *out-of-copyright* group because they could potentially benefit from uncopyrighted issues, while the rest are assigned to the *in-copyright* group for analysis. The research design then estimates the impact of copyright on reuse by comparing the change in content for in-copyright and out-of-copyright baseball pages before and after the Google Books digitization event in a differences-in-differences (DD) framework. In other words, if out-of-copyright baseball pages see a larger increase in information after the digitization event than in-copyright pages after controlling for player and time-dependent fixed effects, we can attribute this difference to be the causal impact of copyright on the reuse of information. An additional level of analysis then uses DD estimates obtained separately from both baseball and basketball players to estimate a "differences-in-differences-in-differences"(DDD) estimate of the impact of copyright on reuse that is robust to many alternate explanations.

The results show that the digitization of *Baseball Digest* had a large positive impact on Wikipedia baseball pages, increasing the number of images by 140% and text by about 50% as compared to basketball pages. However, out-of-copyright pages benefit disproportionately compared to in-copyright pages, suggesting that copyright does indeed prevent reuse in this setting. Out-of-copyright pages receive a boost of about 80% more images, and 140% more citations in OLS specifications. However, such gains are unevenly distributed. The positive impact on reuse is many orders of magnitude larger for images than for text, suggesting that copyright is more effective in preventing the reuse of rich media like images and video than of text. Further, copyright seems to disproportionately harm players in the bottom 50 percentile of the quality distribution, suggesting that players who have few alternate sources of information benefit most from the release of out-of-copyright archival material. Finally, the difference in the amount of information on Wikipedia pages caused by the 1964 copyright experiment seems to have real impacts on Wikipedia's readership– out-of-copyright pages receive a boost in monthly page-views of about 30-80% as compared to in-copyright pages according to fixed-effects and instrumental variables specifications. A back-of-the-envelope calculation (see Appendix A.3) suggests losses to Wikipedia to the tune of about $370,000 annually due to restrictions around the reuse of copyrighted material.

In sum, I find that while digitization projects can help the broader diffusion of knowledge, copyright can have negative impacts on reuse. More broadly, this research shows how copyright policy can strongly influence the rate and direction of follow-on innovation on the internet. The present study joins the nascent empirical literature on copyright as well as related work on innovation, digitization and privacy. Of note is work in the legal literature on copyright that estimates the impact of copyright on availability and compares works produced before and after the US copyright cutoff date of 1923 (Heald, 2007, 2009a; Buccafusco and Heald, 2012). Other work that studies the impact of copyright on prices and access in the market for historical books (Li et al., 2012; Reimers, 2013) is also relevant. This paper adds to this literature by focusing on the causal impact of copyright on *reuse*, by using micro-data on knowledge flows and by separately identifying the interaction between digitization and copyright. This study is also related to the literature on privacy (Goldfarb and Tucker, 2011) which focuses on the role of informational frictions in digital markets. Finally, this paper contributes to a broader literature on knowledge flows (Agrawal et al. (2006), Catalini (2012), Singh and Marx (2013)) and innovation and the role of intellectual property (Murray and Stern (2007), Mowery, Thompson, and Ziedonis (Mowery et al.), Hegde (2011)).

The rest of the paper is organized as follows. Section 2 describes the empirical setting including the *Baseball Digest* experiment and data collection. Section 3 analyzes the impact of the *Baseball Digest* copyright experiment. Section 5 concludes.

2 Empirical Context and Data

2.1 Empirical Context

Google Books and the Baseball Digest Copyright Experiment

Google Books is a Google initiative that has as its objective the digitization of all books ever published and currently offers a catalog of about 30 million books. It is perhaps the most prominent of a number of ongoing digitization projects which include US government efforts to make available digitized records from government transactions and the Library of Congress "National Digital Library" project.

I focus on an event that occurred on 9 December 2008, when Google Books announced that it would make available all past issues of *Baseball Digest*, along with a number of other popular magazines like *Popular Mechanics* and *Ebony*. All published issues since the magazine's founding in 1942 were made available all at one time with consent from the publishers, representing a large increase in the amount of information about the game that was available in digitized format. This is important, not only because *Baseball Digest* is one of the most important references about the game of baseball, but also because of a legal detail that ensures that issues of *Baseball Digest* published before 1964 are in the public domain, while those published after are under copyright.

While it is generally assumed, both in practice and in the literature, that copyright status of works varies only if they were published before 1923 (rendering them in the public domain), additional variation within a single publication can be obtained from the *copyright renewal* requirement. While copyright renewals are no longer necessary, the 1909 Copyright Act that governs works published until 1964 provided for two copyright terms: a 28-year initial term and a 28-year renewal term. However, the renewal term was not automatically granted, and if the renewal application was not filed on time, the work entered the public domain. While copyrights for some movies and books were renewed successfully, because this requirement was not well-known, it *"tripped up many smaller publishers, and a failure to renew caused many works to lapse into the public domain"* (Andrade, 2014). The University of Pennsylvania library's "Copyright Renews for Periodicals" shows that due to a failure to renew copyrights, many other published works during this period have fallen into the public domain including journals like the *American Economic Review*, *Bell System Technical Journal*, *Biometrika* and periodicals like *The Baltimore Sun*, *The Los Angeles Evening Herald*, *The Kansas City Star* etc. Appendix A.1 provides more background on this topic. I conducted a thorough review of the copyright renewals register and found no evidence that *Baseball Digest* had been renewed during this period. Because no renewals were found, it is safe to assume that all issues published before 1964 have fallen into the public domain.

Because it was both digitized by Google Books and failed to renew copyrights, *Baseball Digest* is especially useful for understanding both the impact of copyright on reuse and the role of digitization. Further, the experiment is likely to be economically meaningful given the widespread interest in both the game of baseball and in *Baseball Digest.* Over 45% of all Americans identify as baseball fans, and revenues from the sport of baseball in 2010 were estimated to be approximately 7 billion USD. *Baseball Digest* has provided baseball's vast fan-base with information and news about the game over seven decades since its founding in 1942 (Jones (2007), Brown (2011)).

The Google Books digitization of *Baseball Digest* therefore represents a unique case. The publication date of the periodical (before or after 1964) determines whether a particular issue was under copyright and the date of access (before or after December 2008), determines whether it was digitized. The role of differing digitization and copyright status on the reuse of content is the empirical focus of the paper.

2.2 Data

In order to understand the impact of *Baseball Digest*'s copyright status on reuse, I turn to Wikipedia. There are many reasons why Wikipedia is a natural venue for such analysis. First, Wikipedia is the preeminent source of information on the internet. 56% of typical Google noun searches point to a Wikipedia page as their first result, and 99% point to a Wikipedia entry on the first page (Silverwood-Cope, 2012). Second, Wikipedia is built explicitly on the "No Original Research" rule which requires editors to cite a secondary source for contributions, making the use of magazines like *Baseball Digest* typical on the site. Third, *Baseball Digest* often contains profiles of baseball players and teams, particularly in the form of detailed articles, interviews and player images. Such biographical information forms the foundation of any encyclopedia (Greenstein and Zhu, 2014) and is therefore particularly likely to be reused on Wikipedia. Finally, each revision of a Wikipedia page is archived and publicly accessible. This allowed me to collect repeated panel data on Wikipedia pages both before and after a digital version of *Baseball Digest* was made available. This, in turn, allows me to trace the diffusion of cultural content using micro-data, which would be difficult to do in another setting. While one aim of this paper is to show how Wikipedia can be used to study the diffusion of cultural content, others have used similar data to study collaboration and digital knowledge production (Zhang and Zhu, 2010; Greenstein and Zhu, 2012; Nagaraj et al., 2009; Algan et al., 2013; Gorbatai, 2012; Aaltonen and Seiler, 2013).

The dataset that I build is based on four different sources. First, I used the "Baseball Hall of Fame" voting dataset by Sean Lahman[1] to compile a list of 541 players who have been nominated for election to the Baseball Hall of Fame and who made their debut appearances between 1944 and 1984. I choose these years because they form a window of about 20 years before and after the copyright cutoff year of 1964. The Hall of Fame nomination list allowed me to include players who had finished their careers and who had passed a screening committee judgment, but it also "removes from consideration players of clearly less qualification" (Abbott, 2011). Thus, the nomination list can be said to include only those players who merit encyclopedic inclusion. The dataset also provides biographical details of the players including date of debut and performance details like their number of appearances and the length of their playing career.

In addition to data on prominent baseball players, I also collected similar data on a comparable group of basketball players. While the core of the analysis focuses on analyzing the impact of copyright within the set of baseball players, adding another layer of controls using basketball player data helps strengthen my analysis. This data allows estimation of the baseline impact of digitization, in addition to providing additional controls that can account for overall dynamics on the Wikipedia platform. Basketball is a convenient choice because like baseball, basketball also has an active community of editors on Wikipedia. However, basketball did not experience any digitization event like the availability of *Baseball Digest* through Google Books. For basketball, I obtained data from databasebasketball.com which provides names, years of debut and other performance data for individual players. To identify a set of players comparable to the baseball players, I used the dataset for the Top 1000 players by career minutes played and choose those who

[1] see http://www.seanlahman.com/baseball-archive/statistics/

made their debut between 1944 and 1984.[2]

Having constructed this data, I then manually matched the names of players to their respective pages on Wikipedia. Manual matching helps avoid problems where a player with a common name like "Jackie Robinson" is matched to the Wikipedia page for Jack Robinson the politician, or worse, Jackie Robinson the basketball player. After having completed this matching, for each player page I downloaded archival versions of each player's page as it appeared on December 1 for every year between 2001 and 2013.[3] I then built an automated python parsing utility that allowed me to measure citations to *Baseball Digest* (as measured by references to *Baseball Digest* in the text), the number of images[4] on a page, and number of words of text. Finally, I obtained web traffic data in the form of page-views from `stats.grok.se`. For each player page, I computed average monthly traffic data for every year from 2013 back to 2007, before which traffic data is not available. Additionally I constructed a *quality* metric for each player. For baseball players, *quality* is calculated based on percentile rank in the list of number of all-star appearances.[5] For basketball players the *quality* rank is assigned based on the number of career minutes played. *Quality* is a categorical variable with 4 values, indicating the player's ranking by percentile in their given sport (top 25 percentile, 25-50 percentile, 50-75 percentile and bottom 25 percentile).

Table 1 lists summary statistics for both baseball (Panel A) and basketball (Panel B) players. The first main independent variables of interest are *1(Debut in out-of-copyright year)* which I include in the analysis as $out-of-copy$, which assigns a player to the out-of-copyright group. The second is, *1(Year>2008)* which I include in the analysis as *post* and which indicates that the observation is from a time period after the Google Books digitization event. The main dependent variables are *Citations, Images, Text* and *Traffic.* The data show that baseball Wikipedia pages contain an average of 0.15 citations, 0.56 images and 969 words and that they receive on average 105 page views per month. For basketball players, Wikipedia pages have an average of 0 citations to *Baseball Digest*, 0.19 images and 540 words and receive an average of 123 page views per month.

[2] Basketball Hall of Fame nominations data is not feasible because the voting process does not begin until 1959

[3] December 1 is a convenient date, because it provides measurements in yearly intervals from the digitization event which happened on December 9. 2001 was the year when Wikipedia began and 2013 is the final year of analysis.

[4] I detect images by looking for references to the following file extensions: `jpg,jpeg,gif,svg,tiff,png`

[5] The All-Star game is an annual event that takes place between the "best" players of baseball's two leagues, and therefore provide a good indicator of a player's performance in a given year.

3 Empirical Results

In this section, I investigate the impact of the Google Books digitization on the reuse of information within Wikipedia by comparing baseball player and basketball player Wikipedia pages. I then proceed to understanding the differential impact of this information across the copyright cutoff date of 1964, first within the sample of baseball pages and then using the full sample.

3.1 Impact of Digitization on Reuse

In order to estimate the impact of the digitization of *Baseball Digest* on Wikipedia, I exploit the fact that while baseball pages were at risk of being affected by a digital baseball magazine, there was no comparable digitization event that basketball player pages could have benefited from in this period. Indeed, a search of Google Books, provides no evidence of an equivalent *Basketball Digest* event, making basketball an ideal control sport.

I estimate OLS models with the following specification:

$$Y_{itg} = \alpha + \beta_1 \times post_t \times baseball_g + \gamma_i + \delta_t + \epsilon_{itg}$$

where γ_i and δ_t represent player and time fixed effects respectively for player i, in sport g and year t. The variable $baseball_g$ is an indicator variable that equals one for baseball players and zero for basketball players, while $post_t$ is an indicator variable that equals one if the observation is from a year after 2008. The player fixed effects control for level differences in quality or interest across different players that might influence the quality of their page, while time fixed effects control for changes in general interest in Wikipedia over time. The coefficient β_1 on the variable of interest, $post_t \times baseball_g$, estimates the differential impact on baseball Wikipedia pages relative to basketball ones, after a digitized version of *Baseball Digest* became available. Estimates are presented from OLS models and standard errors are clustered at player level; this allows for the potential serial correlation in the error terms at the level of players and uses repeated observations on pages of the same player to estimate standard errors robust to this problem. This is similar to the commonly used method of clustering standard errors at the state level in difference-in-differences analyses with individual–state–time level observations (Bertrand et al., 2004). Standard errors are clustered in a similar way for all the specifications reported in this study.

Table 3 presents estimates from this specification. Column (1) presents the estimates with player and year fixed effects. Columns (2) adds sixty-five indicator[6] variables at the *decade* × *year* level, that help model different time trends for each player-cohort depending on the decade of their debut (from the 1940s upto the 1980s). Similarly, Column (3) adds fifty-two *quality* × *year* indicator variables[7] at the decade-year level, that help model different time trends for each player-cohort depending on their overall quality rank in their sport, relaxing the assumption that player-pages of different quality levels evolve in a similar manner. While player effects control for level differences in player quality, Columns (2) and (3) help address the concern that older or better players might coincidentally have a more robust growth over time confounding the impact of digitization (or copyright) in this setting.

The estimates suggest that baseball pages benefited significantly more from the digitization event than basketball pages, in terms of the number of citations, images and amount of text. Dividing the coefficients by the mean values suggests that citations experienced an increase of about 500%, and increase in the number of images by about 140% and an increase in the amount of text by about 50%. The estimates in Columns (2) and (3) are similar, albeit smaller in magnitude, suggesting that differing evolution of player information from different decades or quality cohorts is not driving the baseline results.

3.2 Impact of Copyright on Reuse

Having established that Wikipedia pages benefit from digitization in terms of content, I now turn to analyzing the central question of this study: does copyright affect the reuse of digitized content on Wikipedia?

3.2.1 Cross-Sectional Estimates and Graphical Analysis

The basic comparison between out-of-copyright (debut before 1964) and in-copyright (debut after 1964) player pages in 2013 can be presented in a simple cross-tabulation. Table 2 compares citations to *Baseball Digest*, number of images, the amount of text, and the average monthly traffic to out-of-copyright and in-copyright baseball pages in Panel A. These data suggest that, as of December 2013, in-copyright player pages have a lower amount of citations, images, text and traffic as compared to out-of-copyright pages. For example, out-of-copyright pages have about 0.85 more images (an

[6] 13 years × 5 decades = 65

[7] 13 years × 4 quality dummy variables = 52

82% difference) and about 96 more visitors per month (a 103% difference) than in-copyright pages. The corresponding comparison between out-of-copyright and in-copyright basketball players reveals no statistically significant differences between the two groups, suggesting that differences in information and traffic across the 1964 cutoff are specifically related to baseball pages on Wikipedia.

The main concern in attributing these present day differences to the copyright status of *Baseball Digest* is that older players might have been of inherently higher interest to the Wikipedia community than newer players, and I may be attributing these differences to the 1964 copyright cutoff. Figure 2 provides some preliminary but illustrative evidence highlighting why this is unlikely to be the case. In order to generate this chart, the change in the number of images right before the digitization event (2008) and at the latest time period (2013) was calculated for each player. Then I assigned each player to a "debut cohort." That is, all players who made their debut in a given year were grouped together and the mean change in the number of images after digitization was calculated. This mean change was adjusted to account for the fact that even though some players might have made their debut before 1964, they played the majority of their career in the in-copyright period; correspondingly, I scaled the gain in images to account for this difference.[8] This mean change is plotted in dark-gray bars for players in the out-of-copyright group, while the bars in light-gray represent players in the in-copyright group. Panel A plots the data for change in images after digitization for baseball players in the out-of-copyright (debut before 1964) and in-copyright (debut after 1964) groups, while Panel B plots similar data for basketball players.

The resulting plot indicates that baseball players in the out-of-copyright group have a consistently larger increase in the number of images after digitization as compared to players in the in-copyright group. As an illustration, all baseball players in the data who made their debut in 1957, gained about 2.59 images on average while those making their debut in 1972 gained only about 0.68 images. Further, after the 1964 cutoff, this gain seems to decrease in a discontinuous manner, suggesting the importance of the 1964 copyright cutoff in driving these differences and providing less support for the theory that these changes are driven simply by a preference for older players over newer ones. Therefore, this plot suggests that the simple cross-sectional results highlighted in Table 2 are likely to hold even when considering changes in the amount of information before and after the copyright cutoff year of 1964. Panel B indicates that for basketball players,

[8] The scaling is performed by debut cohort and according to this formula – $\Delta Img_{scaled} = \Delta Img \times \frac{FinalYear-DebutYear}{1964-DebutYear}$. The plot looks largely similar even without this rescaling, although player groups who make their debuts roughly around the 1962-1964 window do see a smaller gain in the number of images.

not only is there a smaller growth in the amount of information over time, but there are also little systematic differences around the 1964 cutoff. This confirms the intuition that the copyright cutoff of 1964 is relevant only for baseball pages.

3.2.2 Within-Baseball Estimates

Overall, both Table 2 and Figure 2 seem to indicate that out-of-copyright baseball players seem to have benefited disproportionately as compared to players in the in-copyright group after digitization. This section tests this idea systematically in a regression framework and forms the baseline specification for this study.

Specifically, I estimate:

$$Y_{it} = \alpha + \beta_1 \times post_t \times out - of - copy_i + \gamma_i + \delta_t + \epsilon_{it}$$

where γ_i and δ_t represents player and time fixed effects respectively for player i and year t. $out - of - copy_i$ is an indicator variable that equals one if a player makes his debut before 1964 and indicator variable $post_t$ equals one if the observation is from any year after 2008. The coefficient β_1 on the variable of interest $post_t \times out - of - copy_i$ estimates the differential impact on out-of-copyright Wikipedia pages as compared to in-copyright ones after the baseball digest digitization event.

Table 4 presents estimates from OLS models, with standard errors clustered at the player level. As before, Column (1) presents results from the baseline specification with year and player fixed effects, while Columns (2) and (3) include $decade \times year$ and $quality \times year$ fixed effects that allow for a more flexible time trend that differs across player-debut decades and quality cohorts respectively. The estimates suggest that out-of-copyright players have a greater increase in information than in-copyright players post 2008, even after controlling for player fixed effects and year fixed effects, thus confirming the intuition of Figure 2. For brevity, I focus on interpreting the magnitude of coefficients in Column (1). The estimate in Panel A suggests that citations in out-of-copyright pages grew by about 0.215 (a gain of 140%), Panel B suggests a gain in images of about 0.441 (or 80%) per page. Panel C suggests a gain of about 317 words (or 30%) per page. Therefore, the baseline estimates reveal that the copyright cutoff of 1964 had a meaningful impact on the reuse of content from *Baseball Digest*.

A second finding from Table 4 is that the negative impact of copyright is larger for the reuse of images than for the reuse of text, a growth of 80% for images as compared to 30% for text. Further analysis also indicates that the impact of copyright on the reuse of images is robust to different specifications, while the growth is text is often indistinguishable from zero in many models. Interviews with participants suggest that this is likely because reusing textual information while circumventing copyright is possible through paraphrasing and rewording content, which is not the case with media like images, video, or music. The fact that copyright only prohibits the verbatim copying of information (while allowing paraphrasing) has real implications for Wikipedia – the estimates are much smaller for the reuse of text, while they are quite large for the reuse of images. This result highlights a nuance of the impact of copyright on reuse that is missing from the policy discussion on this topic, i.e. that copyright provides differential protection depending on the *medium* in which the information is expressed.

Finally, in the appendix (Table A.2), I estimate a similar specification using log models (with $Log(Y_i + 1)$ as the dependent variable) as a robustness check. These models also show that out-of-copyright pages benefit disproportionately than in-copyright pages in terms of citations and the reuse of images. As stated, the coefficient in Panel C that estimates the impact of copyright on text becomes smaller and indistinguishable from zero. The coefficients in Panels A and B retain their statistical significance, however they are significantly smaller in size – the impact on citations is estimated to be in the range of 8%, while the impact on images is on the order of 13%.

3.2.3 Triple Difference Estimates

Having established that digitization affects baseball pages positively and that copyright seems to hurt the extent of reuse for baseball players in my baseline estimates, I use data from the basketball players to provide an additional check to my results. The discontinuous cutoff around 1964 that we see in Figure 2 and the fact that controlling for either $decade \times year$ or $quality \times year$ does not change the results substantially provides some reassurance that results are not being driven by differences in the evolution of editing activity for different player debut or quality cohorts. However, there is still the possibility that Wikipedia-wide trends might be driving the differences. For example, a site-wide campaign to update information from older pages around 2008 could be affecting the results. As a final check, in this section, I compare the evolution of Wikipedia pages for baseball player and basketball players using graphical analysis and a differences-in-differences-in-differences

(DDD) regression framework.

Graphical Analysis

First, to explore the timing of estimated effects and to see how they differ between baseball and basketball pages, Figure 3 presents graphical versions of the following event study specification separately for baseball (Panel A) and basketball (Panel B) players where t represents calendar year.

$$Y_i t = \alpha + \gamma_i + \delta_t + \Sigma_t \cdot \beta_t \cdot out - of - copy_i \times 1(t) + \epsilon_{it}$$

Time varying coefficients (Figure 3 Panel A) reveal no discernible evidence in the increase in images for players in the out-of-copyright group before the digitization event. This is reassuring, because it shows that there are no pre-existing trend differences in the evolution of out-of-copyright and in-copyright pages before the digitization of *Baseball Digest* which suggests that the differences-in-differences regression in Table 4 is well-specified. Estimates also show that coefficients are close to zero before the digitization event and start increasing around 2010, suggesting a lag of about two years before images are reused on Wikipedia. Copyright has a similar negative impact on reuse for text (Panel B) and citations (Panel C), again suggesting no differences in the pre-trends between in-copyright and out-of-copyright groups and an increase in reuse post-digitization.

As an additional check, panels A and B also plot estimates from a regression using the sample of basketball players. These “placebo” charts show that there are very little systematic differences between the two groups before and after 2008 for basketball players. Basketball pages never cite *Baseball Digest*, so Panel C is not estimated for that group.

Regression Estimates

Having established that there were no pre-existing differences between the evolution of out-of-copyright and in-copyright baseball pages on Wikipedia, and having compared these changes to basketball pages, I now to turn to formally estimating a DDD regression that uses my full sample of data. Specifically, I estimate

$Y_{itg} = \alpha + \beta_1 \times out - of - copy_i \times post_t \times baseball_g + \beta_2 \times out - of - copy_i \times post_t + \beta_3 \times baseball_g \times post_t + \gamma_i + \delta_t + \epsilon_{igt}$

for player i in year t and in sport g where γ_i and δ_t represents player and time fixed effects respectively. As before $baseball_g$ is an indicator variable that equals one for all baseball players, $out-of-copy_i$ equals one for all players in the out-of-copyright group and $post_t$ equals one for all years after 2008. β_1 is the coefficient of interest, indicating the difference-in-differences-in-differences estimate of the impact of copyright on out-of-copyright baseball players in the *post* period.

The results (Table 5) indicate that the reuse of images on Wikipedia pages does seem to be linked causally to the 1964 copyright cutoff even after adding data from basketball players and thereby controlling for Wikipedia-wide trends. The coefficients are largely similar as compared to the DD specification, indicating a gain of about 0.215 citations and 0.375 images for out-of-copy baseball players. The coefficient on the reuse of text is much smaller as compared to the DD specification and loses its significance, indicating again that the impact of copyright is mainly concentrated in the reuse of images rather than text. Similar to Section 3.2.2, I estimate a version of Table 5 using log transformed dependent variables in the Appendix Table A.3. Again, the coefficients are smaller in magnitude (indicating a gain in the range of 8% for citations and 11% for images) but retain their statistical significance. The coefficient on the reuse of text is indistinguishable from zero.

3.2.4 Further Robustness Checks

So far, I have assembled five pieces of evidence that link the variation in copyright on *Baseball Digest* to the differences in the amount of information on baseball Wikipedia pages – the simple cross section analysis (Table 2), the chart plotting mean differences across debut cohorts (Figure 2), DD regressions with player fixed effects, $decade \times year$ and $quality \times year$ fixed effects (Table 4), a plot of time-varying coefficients (Figure 3), and, finally triple-differences estimates using data from both baseball and basketball pages (Table 5). Together, this evidence strongly suggests that the 1964 copyright cutoff had a negative impact on the reuse of information from *Baseball Digest* on Wikipedia after controlling for a number of alternative explanations. Such explanations include differing time trends for different player vintages and quality levels and differences in pre-trends between out-of-copy and in-copy groups (which were shown not to exist).

In Appendix Table A.4, I present four additional models that further investigate the robustness of the baseline estimates presented in Table 4. In the main specifications, out-of-copyright group is defined as all players who made their debut before 1964, and therefore includes some players whose career spanned the 1964 copyright cutoff. Even though this definition is conservative (i.e. it classifies players who possibly could not benefit from out-of-copyright content in the out-of-copyright group and never mis-classifies players in the other direction), Column (1) estimates another model that drops such players who made their debut before 1964 but retired after 1964 from the analysis. The dropped observations cause the coefficient on citations in Panel A to lose its significance, but the coefficient in Panel B increases in size, showing that the classification scheme I use is conservative. Column (2) also tries to deal with this issue of mis-classification, but instead of dropping players, it uses the date that they first made an appearance in an all-star game as the classification year. The idea behind this choice is that a player is likely to be featured in *Baseball Digest* for the first time in the year when he first makes an all-star team, and that this could be a better measure of a player's likelihood of benefiting from from an out-of-copyright magazine. Reassuringly, the estimates in Panel A and B remain significant under this condition, and magnitudes increase in size.

Another concern is that a few extremely well-known players including Yogi Berra and Hank Aaron, made their debuts in the 1950s, and they could be biasing the results. To deal with this issue, I dropped all players who made more than 15 all-star appearances over their career and re-estimated the models. Column (3) shows that dropping these players does not influence the estimates significantly, which confirms the general trend in the data to not be driven by outliers. Finally, in Column (4), I estimate the models using alternate definitions for the dependent variable. Instead of using the count of citations and images, I use an indicator variable that is equal to one if the page has any citations and images at all in Panels A and B. For Panel C, I replaced the count of words as an indicator of text, to a new variable, the *size* of the text matter on a page in kilobytes. Again, the coefficients are significant and positive, although the magnitudes are not comparable due to the different variable definitions.

These robustness checks provide reassurance that the baseline results are robust to alternate treatment definitions, to alternate outcome definitions and to excluding outliers.

3.3 Which Players Benefited from the 1964 Copyright Experiment?

In the analysis so far, I have established that baseball pages benefited from the digitization of *Baseball Digest* and that this impact was largely concentrated among out-of-copyright pages and for the reuse of images.

In this section, I analyze the heterogeneous impact of the 1964 copyright experiment on baseball pages. Specifically, I address what kinds of players benefited from the availability of out-of-copyright material on *Baseball Digest*? I hypothesize that the 1964 copyright cutoff benefited the lower quality players among my sample than it did the high quality players. Information about higher quality players is often available through multiple channels, and *Baseball Digest* is less likely to be uniquely positioned to have such content.[9]

In order to examine the heterogeneous impact of the copyright experiment on Wikipedia pages, I estimate the following specification within the set of baseball players:

$$Y_{it} = \alpha + \beta_1 \times post_t \times out - of - copy_i + \sum_{m=2}^{4} \beta_m \times post_t \times 1(quality_i = m) +$$

$$\sum_{n=2}^{4} \hat{\beta}_n \times post_t \times out - of - copy_i \times 1(quality_i = n) + \gamma_i + \delta_t + \epsilon_{it}$$

where γ_i and δ_t indicate player and time fixed effects respectively. The key indicator variable, $quality_i$ is defined in Section 2.2 and is a categorical variable that indicates the percentile "quality" rank of a player as a number between 1 and 4 (top 25 percentile, 25-50 percentile, 50-75 percentile and bottom 25 percentile). The key coefficients of interest $\hat{\beta}_n$ estimate the difference between the $out - of - copy_i \times post_t$ coefficient and $out - of - copy_i \times post_t \times 1(quality = n)$ coefficient for each quality percentile n. In other words, $\hat{\beta}_n$ provides estimates of the differences in the impact of copyright on reuse for players of different quality levels.

Figure 5 plots these coefficients separately for each quality percentile.[10] Panel A validates the hypothesis that the impact of copyright on the reuse of images is larger for the players of lower quality than for players of higher quality. The estimate on the $\hat{\beta}_{n=3}$ coefficient is 0.71 and $\hat{\beta}_{n=4}$ is 0.54 while the coefficients on $\hat{\beta}_{n=1}$ and $\hat{\beta}_{n=2}$ are 0.1 and 0.01 respectively. Panel B, which estimates the impact of copyright on traffic to affected pages, also shows a similar pattern, though the impact on the 50-75 percentile is more imprecisely estimated.

Overall, the evidence in Figure 5 is consistent with the theory that out-of-copyright material is most beneficial to players who are notable enough to be included in the encyclopedia, but not famous enough to have substitute information about them available in other sources. This analysis

[9] For example, famous personalities make appearances in public events and many Wikipedia images are often taken by photographers present at such occasions.

[10] I plot the coefficient on $out - of - copy \times post_t$ for quality=1 and add this estimate to coefficients for other quality levels to compute marginal effects

suggests that an important channel through which the digitization of *Baseball Digest* proved useful to Wikipedia was through the unlocking of unique material about *famous-but-not-superstar* players on Wikipedia.

4 The Impact of Copyright on Traffic

Having established that copyright harmed the reuse of information in Section 3.2, I now turn to investigate whether the differences in the amount of information on in-copyright and out-of-copyright pages have a meaningful impact on internet traffic to Wikipedia pages. Traffic information is calculated as a monthly average for years 2007-2013 (data is not available before this period) and is recorded at the player-page level.

While the cross-sectional results in Table 2 have already indicated that in-copyright baseball pages have lower average monthly traffic as compared to out-of-copyright pages in 2013, these differences could be driven by differences in player popularity across different cohorts. In order to establish the causal impact of copyright on traffic, I use a regression framework that includes player and time effects to account for systematic differences between players.

Table 6 reports estimates from such an analysis. Columns (1) and (2) mimic the specification used in Table 4 in a differences-in-differences framework, while Columns (3) and (4) estimate triple differences models similar to Table 5 with traffic as the dependent variable. The estimates in Column (1) indicate that on average, out-of-copyright pages receive a boost of about 31 hits per month after controlling for player and year fixed effects. The coefficient reduces slightly when $quality \times year$ fixed effects are included. Against a mean of about 105 page-views per month, this represents an increase of about 29%. When basketball data is added, this coefficient increases to 48, representing an increase of about 42%, a slight larger estimate as compared to the DD specification. In the appendix, I examine the robustness of these estimates to log models. Columns (3) and (4) of Table A.5 estimate the impact of copyright on traffic to be about 88.8%, or twice as large as the OLS estimates. However, a conservative estimate of the impact of copyright on traffic to affected pages would be to boost page-views in the order of 30%, a significant difference.

Is the Increase in Traffic Causally Related to an Increase in Images?

So far, I have established that out-of-copyright pages tend to benefit in terms of information (especially images) and traffic after the digitization event. This section tests one potential mechanism driving this effect, i.e. it tests the hypothesis that in-copyright pages have lower increases in traffic after digitization because a lower level of images makes such pages less valuable, making it less likely than a user will access the page. In other words, a lower number of images is the channel that causes a drop in traffic to out-of-copyright pages as compared to in-copyright pages. Anecdotal evidence suggests that image search engines like Google Images that drive substantial traffic to Wikipedia pages, and search engines that list "thumbnail" images next to search results (thereby increasing the their probability of being clicked on) could be behind such an effect.

I perform three tests to investigate the possible link between the number of images and level of traffic. First, Figure 4 plots the change in traffic on the vertical axis against the change in the number of images for Hall of Fame players (Panel A) and all players (Panel B) on the horizontal axis. If traffic increases with the number of images these variables are expected to be positively correlated. Figure 4 shows that the correlation coefficient between the change in traffic and change in the number of images is 0.659 for Hall of Famers and 0.459 for the full sample. This indicates that players whose pages are associated with a larger change in images are also associated with a larger increase in internet traffic. For example, Hank Aaron's page (debut 1954) gains 4 images over this period and is associated with an increase in over 260 hits per month.

While this evidence is suggestive, in order to control for player-level differences, I run a simple OLS regression at the page level, where I regress the level of traffic on the number of images for baseball player pages and player-level controls. Table 7 Column (1) presents the estimate from this regression which indicates that an increase of one image is associated with an increase in about 122 page-views. However, when player level fixed effects are added this estimate decreases to 44.38 page-views per image, suggesting that player level differences not captured by the controls are driving an important part of this effect. While player fixed effects are useful, a problem persists in that images are possibly added to player pages when there is greater interest in the player. For example, a player might win an award that would both increase traffic and increase the probability that an additional image is added to his page. Therefore, naive OLS estimates of the impact of images on traffic, even with player fixed effects, could be biased upwards if the timing of the addition of new images is endogenously affected by the level of traffic.

In the final step of the analysis, I use instrumental variables estimation to tackle this problem. Specifically, I instrument for the number of images in a given year with a dummy for $out-of-copy_i \times post_t$. We expect this instrument to be correlated with the number of images (because images were exogenously more likely to be available for out-of-copyright pages in the *post* period) but the timing of digitization is independent of the level of traffic to baseball pages on Wikipedia. Table 7 shows first stage estimates indicating that the chosen instrument is strongly correlated with the number of images; the instrument seems to explain a large portion of the variation in the number of images.[11] The IV estimates are presented in Columns (5) and (6) – and are estimated to increase traffic in the range of 88-113 page-views per month. This increase in the estimated impact of an image on traffic in the IV specifications, suggests that simultaneity of image addition and increase is unlikely to be a problem in this setting. In fact, in my interviews, editors indicated that they discovered *Baseball Digest* as a source of information and would turn to it in their routine editing activity. Therefore, image addition seems to have little to do with exogenous increases in player popularity that might also drive increased traffic.

Finally, even though the IV estimates are extremely suggestive that an increase in traffic is caused by an increase in images, caution must be exercised in interpreting this estimate because the exclusion restriction is tentative here. Specifically, while the amount of text did not change substantially following digitization, other unmeasured aspects of the page content (like the *quality* of the text) could have be affected by the instrument leading to higher levels of traffic. Taken together, the evidence from these three tests taken together seems to strongly suggest that an important channel through which copyright affects internet traffic to Wikipedia pages is the reduced number of images on a given page caused by copyright.

5 Discussion

> Copyright is out of control. How, even if it's out of control, how does it stifle invention? Anybody can make a movie, and the fact that that movie has a copyright, how does that hurt the Internet, for God's sake?

[11] The F-statistic is reasonably large, but falls slightly below 10 when player fixed effects are added suggesting that this specification could suffer from the problem of weak instruments.

Jack Valenti
Motion Picture Association of America (MPAA)

This paper suggests an empirical framework to answer this question and suggests one mechanism through which copyright might hurt innovation on the internet: prohibiting reuse, particularly within open, community-based innovation projects like Wikipedia.

The main findings of this paper are as follows. First, the digitization of *Baseball Digest* has a positive impact on information of baseball Wikipedia pages. Second, this impact seems to be much larger for out-of-copyright Wikipedia pages than in-copyright pages, which suggests that in this context, copyright harms the reuse of information. Third, the negative impact of copyright seems to be restricted mainly to the reuse of images (rather than the reuse of text) which indicates that the power of copyright to prevent reuse depends on the medium in which the information is expressed. Fourth, the differences in information due to copyright have a real impact on traffic to Wikipedia pages, providing a boost of 30-80% to affected pages. Fifth, out-of-copyright material seems to be most beneficial for players who are notable but not superstars, indicating that digitization of older material might be most useful for content which has few present-day substitutes. Finally, in the appendix, I perform a back-of-the envelope calculation which suggests that a lower bound on the loss to social welfare from copyright is about $372,920 annually for Wikipedia. Appendix A.3 provides the calculations behind this result.

5.1 External Validity

One concern with the results could be lack of external validity. Specifically, one might be concerned that Wikipedia represents an idiosyncratic setting to analyze the impact of copyright on reuse because Wikipedia could be unusually stringent in enforcing copyright which would cause my estimates to be biased upwards. Wikipedia however does not seem to be alone however in enforcing copyright, as a number of other digital platforms, where one might expect reuse of digitized information, also have extensive programs for copyright enforcement. These include YouTube (Seidenberg, 2009), Amazon, all major mobile application stores and even Google's search engine (Dillon Scott, 2011). For instance, Apple's AppStore rejected about a thousand applications in August 2009 because they used copyrighted images and books in their applications.[12] Apple also

[12] See bit.ly/1aXpksj

hosts an online tool where firms can report copyright violation. Meanwhile, Google removed about 26 million links from its search index in October 2013[13] including links that provided access to copyrighted books, music and data. An extensive literature on piracy (Bechtold, 2004) in the entertainment industry has also shown that copyright-related interventions that limit the availability of digitized content are quite common, and are often quite effective (Danaher et al., 2010; Danaher and Smith, 2013).

To further ease concerns about external validity, my study builds on the emerging empirical literature on the effects of copyright that suggests that copyright has a negative effect on access, a first step to reuse. Extant work (Heald, 2007, 2009b; Buccafusco and Heald, 2012) has shown that that works produced before 1923, which are generally in the public domain, are much more accessible today than works produced after 1923. For example, books produced before 1923, are more easily accessible on Amazon and Audible.com and are more likely to be digitized (Brooks, 2005) than those produced afterward. A more recent study in the economics literature (Reimers, 2013) analyzes the market for books in a similar time period and finds that a copyright extension decreases welfare from fiction bestsellers by decreasing variety, thereby causing a decrease in consumer surplus that outweighs the increase in profits. Similarly, a study of the fiction market in the 1820s also shows that an important impact of copyright is likely to be an increase in the price of books and such an increase may reduce access (Li et al., 2012).

In light of the anecdotes from the previous section and recent empirical literature it does seem plausible that the impact of copyright on Wikipedia that is measured in this paper, could generalize to a number of other settings where the reuse of digital information is important. Finally, even if external validity is a concern, given Wikipedia's prominence, the estimates presented represent a significant part of the gains due to innovation in the digital economy.

5.2 Contributions and Managerial Implications

Going beyond the question of external validity, this paper makes a number of contributions to the nascent empirical literature at the intersection of intellectual property, innovation and digitization. To my knowledge, this study estimates, the first causal measurement of copyright on reuse, where the underlying work is a source for creating derivative work. The results are related to another stream of work on the empirical effects of intellectual property on the diffusion of scientific knowl-

[13] https://www.google.com/transparencyreport/removals/copyright/

edge. Some early studies (Murray and Stern, 2007; Murray et al., 2009; Williams, 2013; Furman and Stern, 2011) find a generally negative effect of intellectual property on follow-on use, while more recent evidence seems to be mixed (Mowery, Thompson, and Ziedonis (Mowery et al.), Sampat and Williams (2014), Galasso and Schankerman (2013)). This study provides direct evidence to this literature that the costs of access (i.e. digitization) seem to matter for the impacts of IP on reuse. From a policy point of view, this paper is able to directly address questions that are likely to be important going forward such as: (a) how does the impact of copyright change when works are digitized and access costs are low, and (b) does copyright need to be modified for the digital age?

This study also has implications for managers in knowledge-intensive sectors of the economy. For those in-charge of IP and digitization strategy, this study suggests that copyright can be an effective intellectual property tool to manage digital assets. How effective copyright can be seems to depend on access and the medium in which information is expressed. This is useful because there is often a concern that piracy is so rampant on the internet that tools other than traditional intellectual property (like DRM) are often necessary (Zhang, 2013) to supplement toothless copyright law. Second, for managers who are interested in using user communities like Wikipedia to generate innovation (eg: Boudreau et al. (2011), Franzoni and Sauermann (2014)) or open innovation more broadly (Fleming and Waguespack, 2007), this study suggests that provision of external, uncopyrighted but digitized material can be extremely beneficial. This study finds that knowledge within user communities is often sourced from external sources, and policy measures that affect the availability and legal status of sources can either boost or retard innovative activity within such communities.

Methodologically, this paper provides a number of suggestions for measuring copyright's effects going forward. First, I show how the internet provides a fertile ground for estimating carefully the impacts of copyright on reuse using micro-data. Not only is the internet an important venue where future copyright battles will be waged, but the digital and quasi-permanent nature of digital content allows for the detailed measurement of the creation of new products and services on the internet. Further, in light of the finding that copyright impacts images more than it does text, this research points to the importance of the key difference between patents and copyrights, namely that patents typically protect the underlying idea while copyright protects only the "expression". In other words, the impacts of copyright are likely to vary not simply by the quality of the data, but also by the medium of expression. This distinction is likely to be important going forward.

Limitations and Incentives for Creation

Finally, while this paper does not evaluate the overall welfare consequences of copyright (specifically the impact of copyright on the incentives to create new content is not examined), it is still useful, because it helps estimate what potential incentives copyright needs to provide to creators in order to justify the losses estimated to society.

Even within the case that I study, overall welfare gains from the removal of copyright protection for archival *Baseball Digest* issues could be small if the gains to Wikipedia are offset by significant losses to the publishers of *Baseball Digest*, especially if licensing archival content is a major source of revenue that is hurt by lost copyright protection. I did make a number of reasonable attempts to contact the publishers of *Baseball Digest*[14] in order to investigate the possibility of licensing content for reuse, but my requests were met with no response. This suggests, that in this case, producer surplus from licensing archival material is fairly low.

Even if one believed that copyright is likely to have significant incentive effects on creating new material, this study is useful because copyright issues often arise in policy discussions for works already created. In these cases the argument for extending copyright relies on the assumption that copyright on existing works further the diffusion of information. Such an argument was a feature of the "Mickey Mouse" law of 1998[15]. Further, in the case of archival material, the question of compensating authors does not arise because of the so-called "orphan works" problem, when the authors of material cannot be identified or contacted (Smith et al., 2012). Even under the possibility of copyright creating incentives for creating new material, the estimates help measure welfare losses from retroactive extensions of copyright or from difficulties in locating missing authors.

[14] including emailing them, filling out a contact form and calling their office

[15] Sonny Bono Copyright Term Extension Act, Pub. L. No. 105-298, 112 Stat. 2827 (1998).

References

Aaltonen, A. and S. Seiler (2013). Cumulative knowledge and open source content growth: The case of wikipedia.

Abbott, L. (2011, September). Future baseball hall of fame players who did not appear in a world series.

Agrawal, A., I. Cockburn, and J. McHale (2006). Gone but not forgotten: knowledge flows, labor mobility, and enduring social relationships. *Journal of Economic Geography 6*(5), 571–591.

Algan, Y., Y. Benkler, M. F. Morell, and J. Hergueux (2013). Cooperation in a peer production economy experimental evidence from wikipedia. In *Workshop on Information Systems and Economics, Milan, Italy*, pp. 1–31.

Andrade, B. (2014). Copyright renewal - when it had to happen, or else.

Arora, A., A. Fosfuri, and A. Gambardella (2001). *Markets for technology: The economics of innovation and corporate strategy.* MIT press.

Arrow, K. (1962). Economic welfare and the allocation of resources for invention. In *The rate and direction of inventive activity: Economic and social factors*, pp. 609–626. Nber.

Bechtold, S. (2004). Digital rights management in the united states and europe. *Am. J. Comp. L. 52*, 323.

Benkler, Y. (2006). *The wealth of networks: How social production transforms markets and freedom.* Yale University Press.

Bertrand, M., E. Duflo, and S. Mullainathan (2004, February). How much should we trust differences-in-differences estimates? *The Quarterly Journal of Economics 119*(1), 249–275.

Boudreau, K. J., N. Lacetera, and K. R. Lakhani (2011). Incentives and problem uncertainty in innovation contests: An empirical analysis. *Management Science 57*(5), 843–863.

Brooks, T. (2005). How copyright law affects reissues of historic recordings: A new study. *ARSC Journal*.

Brown, M. (2011, April). MLB revenues grown from $1.4 billion in 1995 to $7 billion in 2010. *Biz of Baseball*.

Buccafusco, C. J. and P. Heald (2012). Do bad things happen when works enter the public domain?: Empirical tests of copyright term extension. *Berkeley Technology Law Journal*.

Catalini, C. (2012). Microgeography and the direction of inventive activity. *Available at SSRN*.

Danaher, B., S. Dhanasobhon, M. D. Smith, and R. Telang (2010). Converting pirates without cannibalizing purchasers: the impact of digital distribution on physical sales and internet piracy. *Marketing Science 29*(6), 1138–1151.

Danaher, B. and M. Smith (2013). Gone in 60 seconds: The impact of the megaupload shutdown on movie sales. *Available at SSRN 2229349*.

Dillon Scott, P. (2011). Google transparency report: UK requests removal of nearly 100,000 items from index.

Fleming, L. and D. M. Waguespack (2007). Brokerage, boundary spanning, and leadership in open innovation communities. *Organization science 18*(2), 165–180.

Franzoni, C. and H. Sauermann (2014). Crowd science: The organization of scientific research in open collaborative projects. *Research Policy 43*(1), 1–20.

Furman, J. and S. Stern (2011). Climbing atop the shoulders of giants: The impact of institutions on cumulative knowledge production. *American Economic Review 101*(5), 1933–63.

Galasso, A. and M. Schankerman (2013). Patents and cumulative innovation: causal evidence from the courts. *Available at SSRN 2247011*.

Gallini, N. and S. Scotchmer (2002). Intellectual property: when is it the best incentive system? In *Innovation Policy and the Economy, Volume 2*, pp. 51–78. MIT Press.

Gans, J. S. (2014, July). Remix rights and negotiations over the use of copy-protected works. SSRN Scholarly Paper ID 2474256, Social Science Research Network, Rochester, NY.

Gans, J. S. and S. Stern (2010). Is there a market for ideas? *Industrial and Corporate Change*, dtq023.

Goldfarb, A. and C. Tucker (2011). Privacy and innovation. Technical report, National Bureau of Economic Research.

Gorbatai, A. (2012, September). Social structure and mechanisms of collective production: Evidence from wikipedia.

Greenstein, S. and F. Zhu (2014). Do experts or collective intelligence write with more bias? evidence from encyclopdia britannica and wikipedia. *Working Paper*.

Greenstein, S. M. (2013, March). Technology: Measuring consumer surplus online | the economist.

Greenstein, S. M. and F. Zhu (2012). Is wikipedia biased? *The American Economic Review 102*(3), 343–348.

Hardin, G. (1968). The tragedy of the commons. *New York*.

Heald, P. (2007). Property rights and the efficient exploitation of copyrighted works: an empirical analysis of public domain and copyrighted fiction best sellers. *UGA Legal Studies Research Paper* (07-003).

Heald, P. (2009a). Testing the over-and under-exploitation hypotheses: Bestselling musical compositions (1913-32) and their use in cinema (1968-2007). *Review of Economic Research on Copyright*.

Heald, P. J. (2009b). Does the song remain the same-an empirical study of bestselling musical compositions (1913-1932) and their use in cinema (1968-2007). *Case W. Res. L. Rev. 60*, 1.

Hegde, D. (2011, February). Tacit knowledge and the structure of license contracts: Evidence from the biomedical industry. SSRN Scholarly Paper ID 1807128, Social Science Research Network, Rochester, NY.

Jones, J. M. (2007, October). Less than half of americans are baseball fans. *Gallup*.

Khanna, D. (2012, November). RSC policy brief: Three myths about copyright law and where to start to fix it.

Kitch, E. W. (1977). Nature and function of the patent system, the. *JL & Econ. 20*, 265.

Landes, W. and R. Posner (2002). Indefinitely renewable copyright. *U Chicago Law & Economics, Olin Working Paper* (154).

Lemley, M. A. (2004). Ex ante versus ex post justifications for intellectual property. *The University of Chicago law review*, 129–149.

Lerner, J., C. Borek, L. R. Christensen, and G. Rafert (2012). Lost in the clouds: The impact of copyright scope on investment in cloud computing ventures. *Harvard Business School Working Paper*.

Lessig, L. (2005, February). *Free Culture: The Nature and Future of Creativity.* Penguin Books.

Li, X., M. MacGarvie, and P. Moser (2012, November). Dead poets' property - the copyright act of 1814 and the price of books in the romantic period. *Working Paper*.

Liebowitz, S. and S. Margolis (2004). Seventeen famous economists weigh in on copyright: The role of theory, empirics, and network effects. *bepress Legal Series*, 397.

Mazzoleni, R. and R. R. Nelson (1998, December). Economic theories about the benefits and costs of patents. *Journal of Economic Issues 32*(4), 1031–1052.

Mowery, D. C., N. C. Thompson, and A. A. Ziedonis. Does university licensing facilitate or restrict the flow of knowledge and research inputs among scientists?

Murray, F., P. Aghion, M. Dewatripont, J. Kolev, and S. Stern (2009). Of mice and academics: Examining the effect of openness on innovation. Technical report, National Bureau of Economic Research.

Murray, F. and S. Stern (2007). Do formal intellectual property rights hinder the free flow of scientific knowledge?: An empirical test of the anti-commons hypothesis. *Journal of Economic Behavior & Organization 63*(4), 648–687.

Nagaraj, A., P. Seetharaman, R. Roy, and A. Dutta (2009, December). Do wiki-pages have parents? an article-level inquiry into wikipedia's inequalities. *Workshop on Information Technology Systems (WITS)*.

Reimers, I. (2013). The effects of intellectual property on the market for existing creative works. *Working Paper*.

Sampat, B. and H. Williams (2014). How do patents affect follow-on innovation? evidence from the human genome. *Working Paper*.

Scotchmer, S. (1991). Standing on the shoulders of giants: cumulative research and the patent law. *The Journal of Economic Perspectives*, 29–41.

Seidenberg, S. (2009). Copyright in the age of YouTube. *ABAJ 95*, 46.

Silverwood-Cope, S. (2012, February). Wikipedia: Page one of google UK for 99% of searches | IP blog: SEO, SMO and web development insights.

Singh, J. and M. Marx (2013). Geographic constraints on knowledge spillovers: Political borders vs. spatial proximity. *Management Science 59*(9), 2056–2078.

Siwek, S. E. (2006). *Copyright Industries in the US Economy: The 2006 Report, prepared for the International Intellectual Property Alliance (IIPA), November 2006.*

Smith, M., R. Telang, and Y. Zhang (2012). Analysis of the potential market for out-of-print eBooks. *Available at SSRN 2141422*.

Varian, H. R. (2006, December). Copyright term extension and orphan works. *Industrial and Corporate Change 15*(6), 965–980.

Watt, R. and R. Towse (2006, December). Copyright protection standards and authors' time allocation. *Industrial and Corporate Change 15*(6), 995–1011.

Williams, H. (2013). Intellectual property rights and innovation: Evidence from the human genome. *Journal of Political Economy*.

Zhang, L. (2013). Intellectual property strategy and the long tail: Evidence from the recorded music industry. *PhD diss., University of Toronto 1.*

Zhang, X. and F. Zhu (2010). Group size and incentives to contribute: A natural experiment at chinese wikipedia. *American Economic Review*, 07–22.

Zittrain, J. (2009). *The future of the internet–and how to stop it.* Yale University Press.

6 Tables and Figures

Table 1. **Summary Statistics**

	Mean	SD	Median	Min	Max	N
Panel A – Baseball						
Number of images	0.56	1.26	0.00	0	19	7033
Number of words of text	969.25	1393.01	582.00	0	16078	7033
Number of citations to Baseball Digest	0.15	0.82	0.00	0	10	7033
Average monthly traffic	105.60	274.79	36.03	0	10088	3787
Year of Wikipedia page version	2007.00	3.74	2007.00	2001	2013	7033
1(Year>2008)	0.38	0.49	0.00	0	1	7033
Debut Year	1966.12	10.19	1966.00	1944	1984	541
Quality Percentile	2.62	1.29	3.00	1	4	541
1(Debut in Out-of-Copyright year)	0.38	0.49	0.00	0	1	541
Panel B – Basketball						
Number of images	0.19	0.69	0.00	0	11	7332
Number of words of text	540.11	1107.85	250.00	0	15516	7332
Number of citations to Baseball Digest	0.00	0.00	0.00	0	0	7332
Average monthly traffic	123.55	709.71	19.25	0	19655	3948
Year of Wikipedia page version	2007.00	3.74	2007.00	2001	2013	7332
1(Year>2008)	0.38	0.49	0.00	0	1	7332
Debut Year	1970.56	9.23	1971.00	1947	1984	564
Quality Percentile	2.10	0.98	2.00	1	4	564
1(Debut in Out-of-Copyright year)	0.20	0.40	0.00	0	1	564

Note: Summary statistics from 7033 baseball (Panel A) and 7332 basketball (Panel B) page-year observations for a total of 14635 page-year level observations between years 2001-2013. *Out-Of-Copyright year* is defined as all years before 1964. Traffic data is only available for years 2007 to 2013. See text for detailed data and variable descriptions.

Table 2. **Reuse outcomes for out-of-copyright and in-copyright pages in 2013**

Panel A : Baseball Players

	(1)out-of-copy $\bar{y}$	**(2)in-copy $\bar{y}$**	**(3)diff**	**(4)p-val**
Number of Citations to Baseball Digest	0.597	0.334	0.263	0.03
Number of Images	1.888	1.036	0.853	0.00
Number of Words of Text	2199.9	1722.8	477.1	0.00
Average Monthly Traffic	190.3	93.35	96.94	0.02

Panel B : Basketball Players

	(1)out-of-copy $\bar{y}$	**(2)in-copy $\bar{y}$**	**(3)diff**	**(4)p-val**
Number of Citations to Baseball Digest	0	0	0	.
Number of Images	0.568	0.411	0.157	0.17
Number of Words of Text	1340.4	1189.4	151.0	0.38
Average Monthly Traffic	134.4	175.8	-41.40	0.66

Note: This table compares reuse outcomes for 2013 versions of out-of-copyright and in-copyright pages. $N = 514$ for baseball (Panel A) and $N = 564$ for basketball (Panel B). Sample in Column (1) includes all pages from out-of-copy players (debut before 1964) and Column (2) includes all pages from in-copy players (debut after 1964). The p-value reported in Column (4) is from a t-test for a difference in mean outcomes across Column (1) and (2). See text for more detailed data and variable descriptions.

Table 3. **Difference-in-Difference Regressions Estimating Impact of Google Books Digitization Event on Baseball Wikipedia Pages**

	(1)	(2)	(3)
Panel A: Citations ($\bar{y}$=0.07)			
baseball X post	0.357	0.329	0.341
	(0.0510)***	(0.0475)***	(0.0549)***
Panel B : Images ($\bar{y}$=0.37)			
baseball X post	0.515	0.479	0.391
	(0.0643)***	(0.0629)***	(0.0636)***
Panel C : Text ($\bar{y}$=750.21)			
baseball X post	394.2	376.0	300.9
	(67.19)***	(67.03)***	(68.95)***
Player FE	Yes	Controls	Yes
Time FE	Year	Decade X Year	Quality X Year
N	14365	14365	14365
Adj R-square	0.406	0.411	0.418
Clusters	1105	1105	1105

*+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01*
Standard errors clustered at player-level shown in parentheses.

Note: Page-year level observations. *baseball*: 0/1, =1 for baseball players and =0 for basketball players. *post*:0/1,=1 if year is greater than 2008. *Decade*, indicator variables for decade of player debut. *Quality*, indicator variables for percentile rank of player in cohort based on sport-specific covariates.

Specification: $Y_{itg} = \alpha + \beta_1 \times post_t \times baseball_g + \gamma_i + \delta_t + \epsilon_{itg}$ where γ_i and δ_t represents player and time fixed effects respectively for player i, in sport g and year t. Main effects for *baseball* and *post* not reported, because

All estimates are from ordinary-least-squares (OLS) models. See text for detailed data and variable descriptions.

Table 4. **Difference-in-Difference Regressions Estimating Impact of 1964 Copyright Experiment on Baseball Wikipedia Pages**

	(1)	(2)	(3)
Panel A: Citations ($\bar{y}$=0.15)			
out-of-copy X post	0.215	0.178	0.201
	(0.112)*	(0.253)	(0.115)*
Panel B : Images ($\bar{y}$=0.56)			
out-of-copy X post	0.441	0.462	0.373
	(0.127)***	(0.171)***	(0.125)***
Panel C : Text ($\bar{y}$=969.25)			
out-of-copy X post	317.8	451.6	275.8
	(117.0)***	(130.1)***	(119.2)**
Player FE	Yes	Controls	Yes
Time FE	Year	Decade X Year	Quality X Year
N	7033	7033	7033
Adj R-square	0.466	0.472	0.476
Clusters	541	541	541

*+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01*
Standard errors clustered at player-level shown in parentheses.

Note: Page-year level observations. *baseball*: 0/1, =1 for baseball players and =0 for basketball players. *post*:0/1,=1 if year is greater than 2008. *Decade*, indicator variables for decade of player debut. *Quality*, indicator variables for percentile rank of player in cohort based on sport-specific covariates.

Specification: $Y_{it} = \alpha + \beta_1 \times post_t \times out - of - copy_i + \gamma_i + \delta_t + \epsilon_{it}$ where γ_i and δ_t represents player and time fixed effects respectively for player i and year t. All estimates are from ordinary-least-squares (OLS) models. See text for detailed data and variable descriptions.

Table 5. **DDD Regressions Estimating Impact of 1964 Copyright Experiment on Baseball Wikipedia Pages using Basketball Pages as 2nd set of Controls**

	(1) Citations	(2) Images	(3) Text
out-of-copy X post X baseball	0.215	0.375	191.8
	(0.112)*	(0.156)**	(180.8)
out-of-copy X post	-1.74e-14	0.0667	126.0
	(5.94e-09)	(0.0915)	(137.9)
baseball X post	0.275	0.361	298.0
	(0.0562)***	(0.0664)***	(66.68)***
Player FE	Yes	Yes	Yes
Time FE	Year	Year	Year
adj. R^2	0.0775	0.197	0.410
N	14365	14365	14365
Clusters	1105	1105	1105

*+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01*
Standard errors clustered at player-level shown in parentheses.

Note: Page-year level observations for full sample. *out-of-copy*: 0/1, =1 for players making debut before 1964. *baseball*:0/1, =1 for baseball players and =0 for basketball players. *post*:0/1,=1 if year is greater than 2008.

Specification: $Y_{itg} = \alpha + \beta_1 \times out - of - copy_i \times post_t \times baseball_g + \beta_2 \times out - of - copy_i \times post_t + \beta_3 \times baseball_g \times post_t + \gamma_i + \delta_t + \epsilon_{igt}$

for player i in year t and in sport g where γ_i and δ_t represents player and time fixed effects respectively. All estimates are from ordinary-least-squares (OLS) models. See text for detailed data and variable descriptions.

Table 6. **Difference-in-Difference Regressions Estimating Impact of 1964 Copyright Experiment on Traffic**

	(1) Traffic	(2) Traffic	(3) Traffic	(4) Traffic
out-of-copy X post X baseball			48.25	48.07
			(28.68)*	(29.44)$^{+}$
out-of-copy X post	31.16	28.12	-17.09	-14.56
	(14.16)**	(14.96)*	(24.95)	(24.39)
baseball X post			-33.54	-42.24
			(18.79)*	(20.53)**
Player FE	Yes	Quality X Year	Yes	Quality X Year
Year FE	Yes	Yes	Yes	Yes
adj. R^2	0.0293	0.0351	0.0290	0.0361
N	3787	3787	7735	7735
Clusters	541	541	1105	1105

*+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01*
Standard errors clustered at player-level shown in parentheses.

Note: Page-year level observations. Samples in Columns (1) and (2) data only for baseball pages, while sample in Column (3) includes full sample. *out-of-copy*: 0/1, =1 for players making debut before 1964. *baseball*: 0/1, =1 for baseball players. *post*: 0/1,=1 if year is greater than 2008. Note that traffic data is not available for years 2001-2006.

All estimates are from ordinary-least-squares (OLS) models. See Table 4 (for Columns 1 and 2) and Table 5 (for Columns 3 and 4) for estimating equations. See text for detailed data and variable descriptions.

Table 7. **Is an Increase in Images Associated With an Increase in Traffic?**

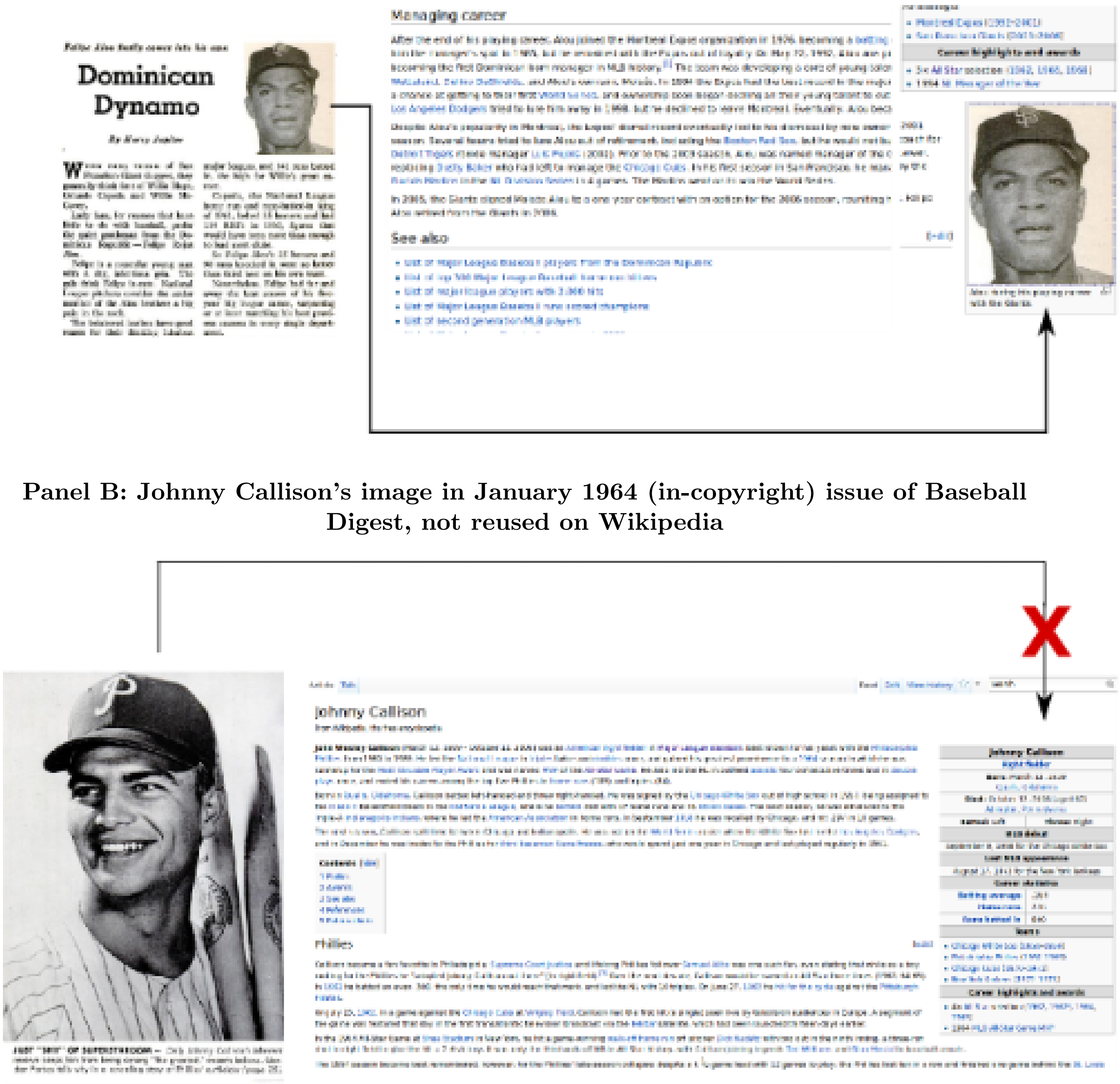
	OLS		First Stage		IV Estimates	
	(1) Traffic	(2) Traffic	(3) Images	(4) Images	(5) Traffic	(6) Traffic
Images	122.0***	44.38*			88.55***	113.8**
	(34.91)	(23.73)			(31.62)	(45.21)
Out-Of-Copy. X Post			0.609***	0.274***		
			(0.154)	(0.0964)		
Controls	Yes	Player FE	Yes	Player FE	Yes	Player FE
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
N	3787	3787	3787	3787	3787	3787
adj. R^2	0.446	0.0737	0.0644	0.190	0.415	-0.0366
F-Stat			15.62	8.07		

*+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01*
Standard errors clustered at player-level shown in parentheses.

Note: Page-year level observations for sample of baseball players. *Images* measures number of images on player-page in a given year. All estimates are from ordinary-least-squares (OLS) models. Specification for Columns (1) is $Y_i = \alpha + \beta_1 \cdot Images_{it} + \delta_t + \epsilon_{it}$, while Column (2) adds individual player dummies.

Columns (3-6) present results from IV estimation, with and without player fixed effects. See text for detailed data and variable descriptions.

Figure 1. **An Illustration of How Copyright Might Affect the Reuse of Information**

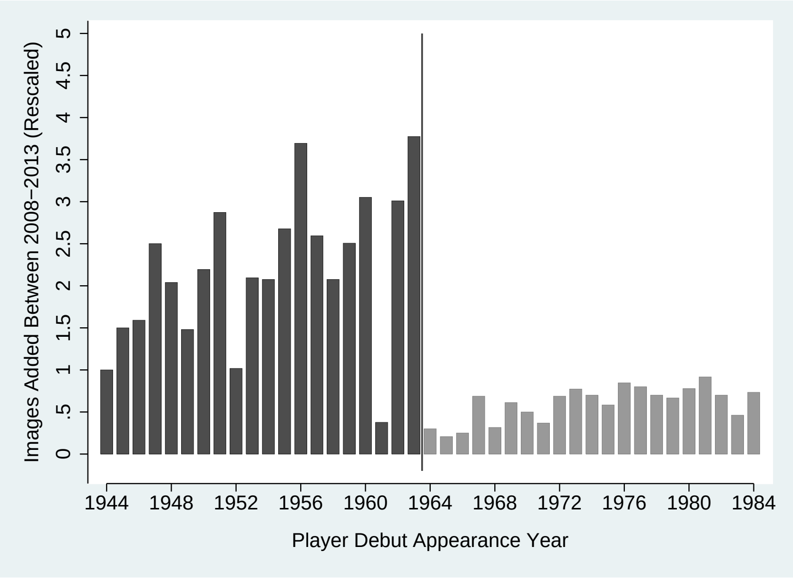


Panel A: Felipe Alou's image in December 1963 (out-of-copyright) issue of Baseball Digest, reused on Wikipedia)

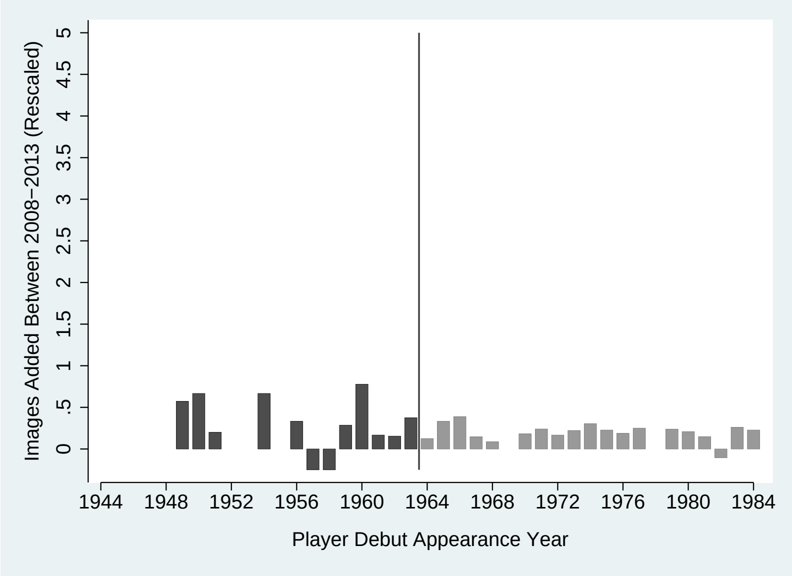
Panel B: Johnny Callison's image in January 1964 (in-copyright) issue of Baseball Digest, not reused on Wikipedia

Figure 2. **Did pre-1964 Players Benefit Disproportionately from Digitization? Mean Images Added (By Debut Cohort) between 2008 and 2013**

Panel A : Baseball Players

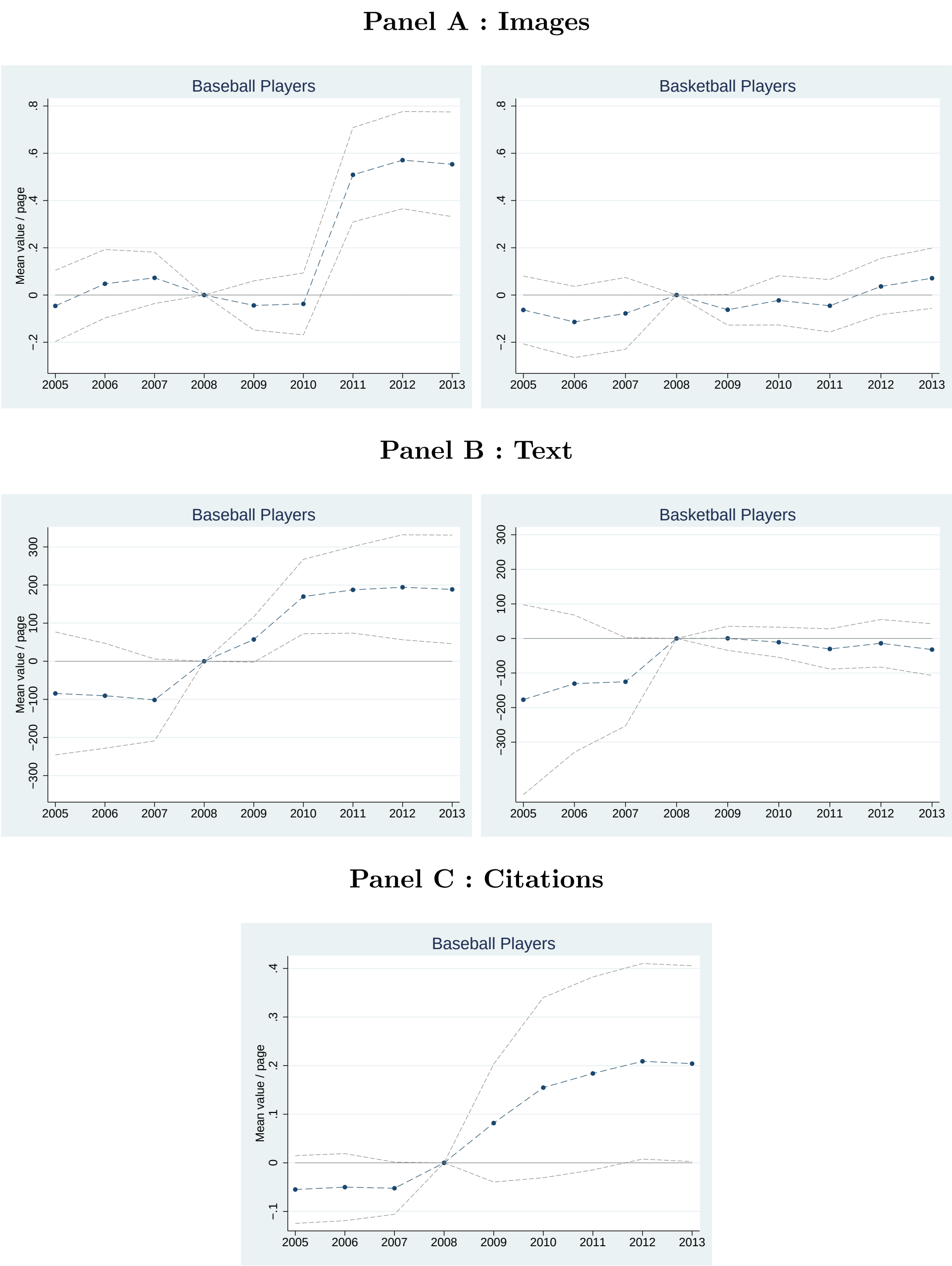


Panel B : Basketball Players



Note: This plot documents the variation in the number of images added between 2008 and 2013 for out-of-copyright (debut before 1964) players and in-copyright (debut after 1964) players. The raw mean is adjusted to account for each player's differing exposure to the copyright rule. See Section 3.2.1 for more detailed data and variable descriptions.

Figure 3. **Time Varying Estimates of the Impact of Copyright on Reuse for Baseball and Basketball Players**



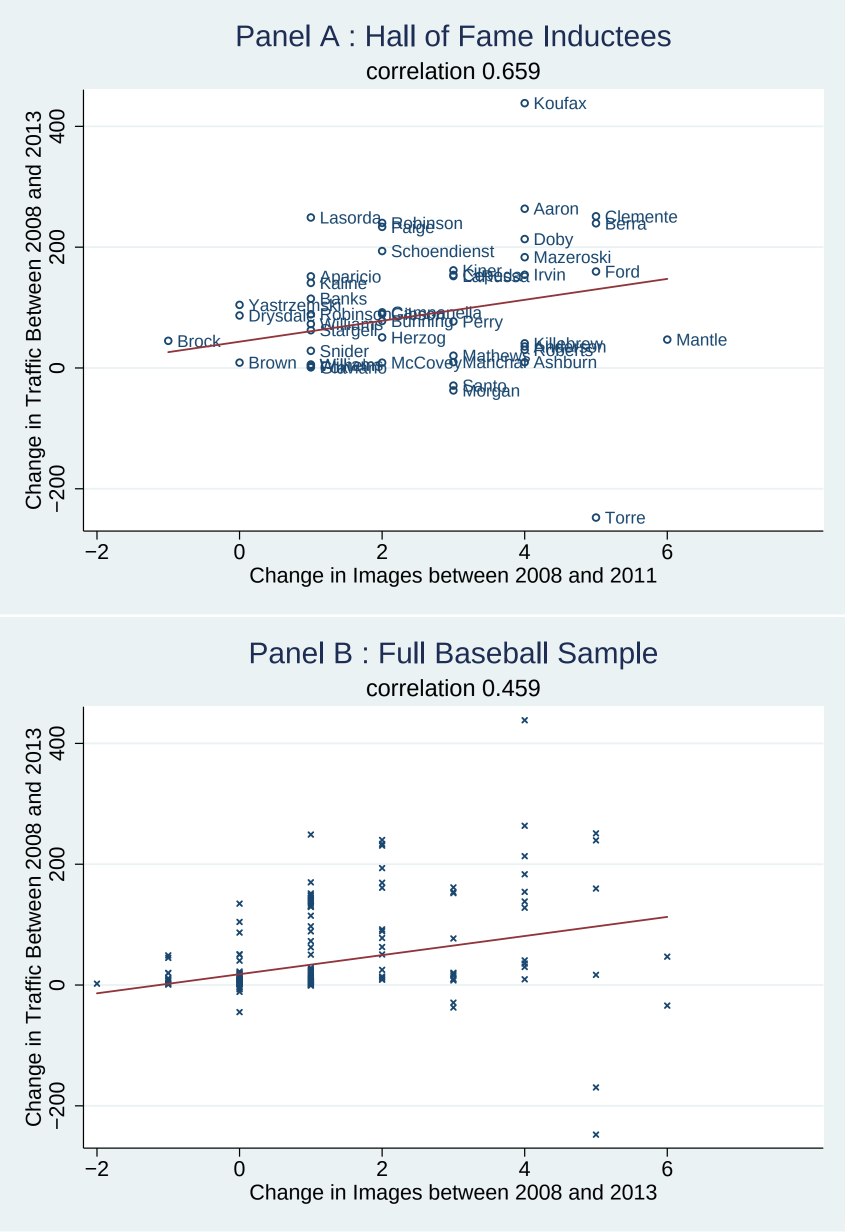
Panel A : Images

Panel B : Text

Panel C : Citations

Note: This figure plots coefficients (and 95 percent confidence intervals) from the event study specifications described in Section 3.2.3. On the x axis is the calendar year and the reference year is 2008, the year of the digitization event. As in Table 4, this specification is based on page-year level observations, the coefficients are estimates from ordinary-least-squares (OLS) models, and standard errors are clustered at the page level. Basketball pages make exactly zero citations in the study period, and therefore Panel C (right) is not estimated for basketball players. See text for more detailed data and variable descriptions.

Figure 4. **Relation between Change in Traffic and Images for Baseball Players**



Panel A : Hall of Fame Inductees

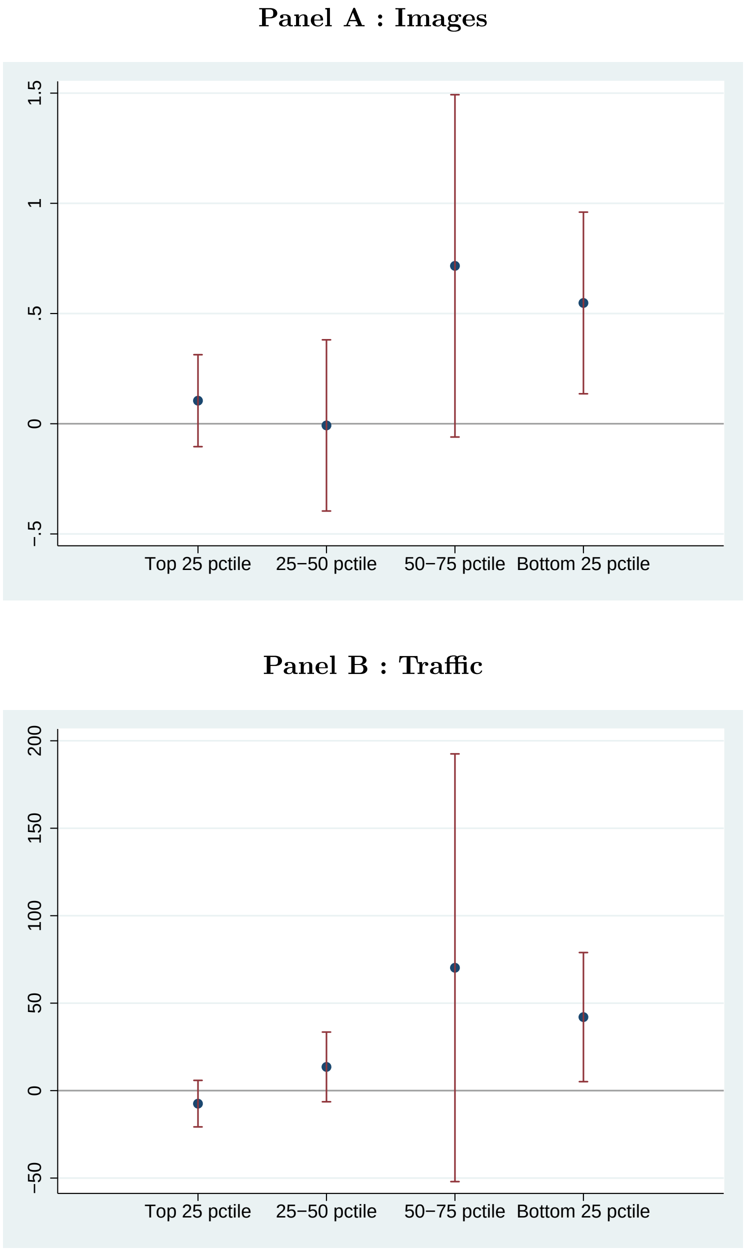
correlation 0.659

Panel B : Full Baseball Sample

correlation 0.459

Note: This plot documents the relationship between the change in images for a given page before and after the digitization of Baseball Digest and associated change in traffic. Panel A, plots the data for all Hall-of-Fame players in the out-of-copyright group and each point is labeled with the player's last name. Panel B is a similar plot for all out-of-copyright baseball players. The correlation coefficient is about 0.659 for Hall of Fame players and 0.459 for the full sample. See text for more detailed data and variable descriptions.

Figure 5. **Heterogeneous Impacts of Copyright on Wikipedia Pages by Player Quality**



Panel A : Images

Panel B : Traffic

Note: This plot documents the differential impact of Baseball Digest copyright cutoff on baseball player pages of different *quality*. For this analysis, the number of times a player was selected to play in an All Star game over his lifetime is measured, and then players are split into 4 different levels of quality based on their percentile rank within the sample of baseball players. See Section 3.3 for detailed data and variable descriptions.

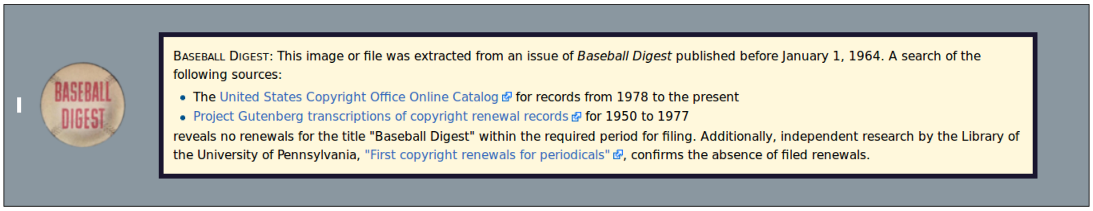
A Appendices

A.1 Appendix A1 : The 1964 Copyright Experiment

Legal opinion about copyright law concerning periodicals (under the Copyright Act of 1909) from University of Pennsylvania Libraries is reproduced below for reference. See http://onlinebooks.library.upenn.edu/cce/firstperiod.html for more details.

> For works that received their copyright before 1978, a renewal had to be filed in the work's 28th year with the Library of Congress Copyright Office for its term of protection to be extended. The need for renewal was eliminated by the Copyright Renewal Act of 1992, but works that had already entered the public domain by non-renewal did not regain copyright protection. Therefore, works published before 1964 that were not renewed are in the public domain. With rare exception (such as very old works first published after 2002) no additional copyrights will expire (thus entering the public domain) until at least 2019 due to changes in the applicable laws.

The following screenshot is from a Wikipedia banner explaining the legal use of an image sourced from Baseball Digest.



BASEBALL DIGEST: This image or file was extracted from an issue of *Baseball Digest* published before January 1, 1964. A search of the following sources:

- The United States Copyright Office Online Catalog for records from 1978 to the present
- Project Gutenberg transcriptions of copyright renewal records for 1950 to 1977

reveals no renewals for the title "Baseball Digest" within the required period for filing. Additionally, independent research by the Library of the University of Pennsylvania, "First copyright renewals for periodicals", confirms the absence of filed renewals.

A screenshot from Wikipedia explaining copyright law pertaining to reuse of material from Baseball Digest

A.2 Appendix A2 : Back-of-the-Envelope Welfare Calculation

This section outlines the methodology through which I arrive at my estimate of about $372,920 annually for a lower bound on the loss to social welfare from copyright.

In order to arrive at this estimate two pieces of data are needed: (a) the approximate value of a page-view to society and (b) estimated page-views lost due to copyright. For piece (a) I used webindetail.com which provides the estimated daily earnings of Wikipedia from potential advertising which would equal to $2.2 million dollars daily for about 400 million daily page-views.[16] This translates into a value to Wikipedia of about $0.0055 per page-view from advertising. For piece (b) results from this study suggest that for every missing image, a Wikipedia page receives about 88 fewer page-views per month (Table 7) and that pages have 0.315 fewer images on average (Table 5) due to copyright. Therefore, a page affected by copyright is expected to lose about $0.152 per month. For the set of 335 pages affected by copyright in this study. This translates into an annual loss of about $612 or a net present value of about $30,600.[17] Assuming that about 5% of all 4.1 million articles on Wikipedia are affected in a similar way, this translates into an annual loss of about $372,920 or a net present value of about $18.69 million. These estimates are economically significant for Wikipedia in the light of estimates of the economic value of Wikipedia itself which is about $43.5 million per year (Greenstein, 2013). Further, these estimates represent only a lower bound on lost surplus because advertising rates capture only the valuation advertisers place on readers and do not calculate value to readers including value of derivative works of Wikipedia pages.

[16] While Wikipedia does not accept advertising, webindetail.com arrives at this estimate based on a comparables analysis based on other similar websites with a comparable user base.

[17] discounted at a rate of 2% per year over a perpetual life term

A.3 Appendix A3 : Robustness Checks

Table A.1. **Log Models: Difference-in-Difference Regressions Estimating Impact of Digitization on Wikipedia**

	(1)	(2)	(3)
Panel A: Citations ($\bar{y}$=0.03)			
baseball X post	0.146	0.136	0.138
	(0.0175)***	(0.0167)***	(0.0186)***
Panel B : Images ($\bar{y}$=0.19)			
baseball X post	0.208	0.201	0.177
	(0.0227)***	(0.0231)***	(0.0229)***
Panel C : Text ($\bar{y}$=4.45)			
baseball X post	-0.461	-0.471	-0.521
	(0.0419)***	(0.0425)***	(0.0427)***
Player FE	Yes	Controls	Yes
Time FE	Year	Decade X Year	Quality X Year
N	14365	14365	14365
Adj R-square	0.853	0.854	0.857
Clusters	1105	1105	1105

*+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01*
Standard errors clustered at player-level shown in parentheses.

Note: Page-year level observations. Column (2) includes 4 indicator variables proxying for a player's "quality" based on sport-specific covariates. *baseball*: 0/1, =1 for baseball players and =0 for basketball players. *Post*:0/1,=1 if year is greater than 2008.

All estimates are from ordinary-least-squares (OLS) models. Specification: $Y_{itg} = \alpha + \beta_1 \times post_t \times baseball_g + \gamma_i + \delta_t + \epsilon_{itg}$ where γ_i and δ_t represents player and year fixed effects respectively for player i, in sport g and year t. See text for detailed data and variable descriptions.

Table A.2. **Log Models: Difference-in-Difference Regressions Estimating Impact of Copyright on Wikipedia**

	(1)	(2)	(3)
Panel A: Citations ($\bar{y}$=0.06)			
out-of-copy X post	0.0819	0.0489	0.0749
	(0.0381)**	(0.0859)	(0.0385)*
Panel B : Images ($\bar{y}$=0.28)			
out-of-copy X post	0.131	0.171	0.113
	(0.0401)***	(0.0785)**	(0.0393)***
Panel C : Text ($\bar{y}$=4.89)			
out-of-copy X post	-0.0119	-0.280	-0.0206
	(0.0682)	(0.146)*	(0.0680)
Player FE	Yes	Controls	Yes
Time FE	Year	Decade X Year	Quality X Year
N	7033	7033	7033
Adj R-square	0.849	0.850	0.856
Clusters	541	541	541

*+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01*
Standard errors clustered at player-level shown in parentheses.

Note: Page-year level observations for baseball sample only. Column (2) includes 4 indicator variables proxying for a player's "quality" based on sport-specific covariates. *Out-of-copy*: 0/1, =1 for players making debut before 1964. *Post*:0/1,=1 if year is greater than 2008.

All estimates are from ordinary-least-squares (OLS) models. Specification: $Y_{it} = \alpha + \beta_1 \times POST_t \times TREAT_i + \gamma_i + \delta_t + \epsilon_{it}$ where γ_i and δ_t represents player and year fixed effects respectively for player i and year t. See text for detailed data and variable descriptions.

Table A.3. **Log Models: DDD Regressions Estimating Impact of Copyright on Wikipedia**

	(1) Citations	(2) Images	(3) Text
out-of-copy X post X baseball	0.0819	0.112	0.0221
	(0.0381)**	(0.0521)**	(0.0918)
out-of-copy X post	-7.44e-15	0.0184	-0.0340
	(6.98e-10)	(0.0333)	(0.0616)
baseball X post	0.115	0.162	-0.463
	(0.0198)***	(0.0264)***	(0.0506)***
Player FE	Yes	Yes	Yes
Time FE	Year	Year	Year
adj. R^2	0.102	0.247	0.853
N	14365	14365	14365
Clusters	1105	1105	1105

*+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01*
Standard errors clustered at player-level shown in parentheses.

Note: Page-year level observations for full sample. Column (2) includes 4 indicator variables proxying for a player's "quality" based on sport-specific covariates. *OutOfCopy*: 0/1, =1 for players making debut before 1964. *Baseball*:0/1, =1 for baseball players and =0 for basketball players. *Post*:0/1,=1 if year is greater than 2008.

All estimates are from ordinary-least-squares (OLS) models. Specification: $Y_{itg} = \alpha + \beta_1 \times OutofCopy_i \times Post_t \times Baseball_g + \beta_2 \times OutofCopy_i \times Post_t + \beta_3 \times Baseball_g \times Post_t + \gamma_i + \delta_t + \epsilon_{igt}$

for player i in year t and in sport g where γ_i and δ_t represents player and year fixed effects respectively. See text for detailed data and variable descriptions.

Table A.4. **Robustness to Sample restrictions, Alternate Variables and Treatment Definition**

	(1)	(2)	(3)	(4)
Panel A: Citations				
out-of-copy X post	0.0375	0.0950	0.0499	0.0581
	(0.0456)	(0.0359)***	(0.0322)$^{+}$	(0.0267)**
Panel B : Images				
out-of-copy X post	0.756	1.021	0.350	0.0624
	(0.298)**	(0.196)***	(0.149)**	(0.0313)**
Panel C : Text				
out-of-copy X post	256.8	700.7	319.8	2238.1
	(260.5)	(186.4)***	(144.1)**	(924.8)**
Player FE	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes
N	4966	7033	5291	5051
Adj R-square	0.484	0.478	0.498	0.444
Clusters	382	541	407	541

*+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01*
Standard errors clustered at player-level shown in parentheses.

Note: This table evaluates the robustness of the impact of copyright on reuse result to different modeling and data assumptions. Page-year level observations for baseball sample only. See Table 4 for specification. All estimates are from ordinary-least-squares (OLS) models.

Column (1) drops all players who overlap the copyright-cutoff year of 1964 and estimates the model using players who retired before 1964 and those who made their debut after 1964. Column (2) uses an alternate defintion of "Out-of-Copyright" using the year of a player's first all star game instead of the debut year for classification. Column (3) drops very well-known players (those who have played 15 all star games or more) and reestimates the model. Column (4) uses alternate dependent variables: Citations and Images are replaced by indicator variables if variable is greater that 0, and traffic is measured by the size of the page in kilobytes. See text for more detailed data and variable descriptions.

Impact of Copyright on Traffic – Alternate Specification

This table provides a robustness check to log models for Table 6. Simple log versions of the models in Table 6 were tried, however a lack of sufficient "pre" data (before 2008) means that the main coefficients were imprecisely estimated, and were not significant at conventional levels. As an alternative, the following table estimates cross sectional regressions that utilize the variance in *copyright exposure* to estimate log models. For each player, *copyright exposure* is defined as amount of their career that they played in the out-of-copyright period, i.e. before 1964. For players who retired before 1964, this index is set to one, for players who made their debuts after 1964 this index is set to zero, while for other players it is calculated as $\frac{1964-DebutYear}{FinalYear-DebutYear}$. Because player debut and retirement years are unlikely to be related to the 1964 copyright cutoff date, this variation provides an additional source of quasi-random variation that can then be used in the cross-section to estimate the impact of copyright on internet traffic, and that helps alleviate the problem of missing traffic data for years before 2007. Columns (1) and (2) show the impact of the Copyright Exposure variable on Images, while Columns (3) and (4) estimate the effect for traffic. Coefficients are roughly the same order of magnitude as with the difference-in-difference specifications.

Table A.5. **Log: Impact of Copyright on Images and Traffic: Robustness with "Out-of-copyright" exposure index**

	(1) Diff. Img	(2) Log Diff. Img.	(3) Diff. Traf	(4) Log Diff. Traf
Out-of-copy Exposure	1.313	0.555	133.8	0.888
	(0.238)***	(0.0706)***	(93.48)	(0.223)***
Constant	0.556	0.319	23.14	2.298
	(0.0484)***	(0.0236)***	(6.671)***	(0.0693)***
Observations	541	541	541	541
Adjusted R^2	0.112	0.121	0.018	0.038

*+:p<0.15; *:p<0.10; **:p<0.05; ***:p<0.01*
Robust standard errors shown in parentheses.

Note: Page-year level observations. Sample includes all baseball pages in 2013. The specification is $Y_i = \alpha + \beta \times out-of-copyindex + \epsilon_i$. All estimates are from ordinary-least-squares (OLS) models, and columns (2) and (4) use $Log(1+Y)$ as the dependent variable.