How Scholarly Is Google Scholar? A Comparison to Library Databases

Jared L. Howland, Thomas C. Wright, Rebecca A. Boughan, and Brian C. Roberts

Google Scholar was released as a beta product in November of 2004. Since then, Google Scholar has been scrutinized and questioned by many in academia and the library field. Our objectives in undertaking this study were to determine how scholarly Google Scholar is in comparison with traditional library resources and to determine if the scholarliness of materials found in Google Scholar varies across disciplines. We found that Google Scholar is, on average, 17.6 percent more scholarly than materials found only in library databases and that there is no statistically significant difference between the scholarliness of materials found in Google Scholar across disciplines.

oogle Scholar was introduced to the world in November of 2004 as a beta product. It has been embraced by students, scholars, and librarians alike. However, Google Scholar has received criticism regarding the breadth and scope of available content. We undertook this study to answer two questions regarding these common criticisms: (1) Are Google Scholar result sets more or less scholarly than licensed library database result sets? and (2) Does the scholarliness of Google Scholar vary across disciplines?

Literature Review

Google Scholar, which is still branded as a beta version, has not only become a common fixture in library literature but is also becoming ubiquitous in information-seeking behavior of users. Google Scholar was initially met with curiosity and skepticism.[1] This was followed by a period of systematic study.[2] More recently, there has been optimism about Google Scholar's potential to move us toward Kilgour's goal of 100 percent availability of information.[3] Librarians now find themselves acknowledging users' preferences for one-stop information shopping by giving Google Scholar ever-increasing visibility on their Web pages.[4] Even as librarians begin to promote Google Scholar, the debate continues within the information community as to the advisability of guiding users to this tool. The view of critics like Péter Jascó, who use terms such as "shallowness" and "artificial unintel-

Jared L. Howland is Electronic Resources Librarian, Thomas C. Wright is Collection Development Coordinator, Rebecca A. Boughan is Electronic Resources Assistant, and Brian C. Roberts is Process Improvement Specialist, all at Brigham Young University. E-mail: jared_howland@byu.edu, tom_wright@byu.edu, rebecca_boughan@byu.edu, brian_roberts@byu.edu.

ligence" to describe the program,[5] seems to be giving way to a landscape where respected publishers (like Cambridge) and platforms (for instance, JSTOR) are now offering links out to Google Scholar for more citations.

Early studies of Google Scholar tried to match citations "hit to hit" in comparison with traditional library databases. Jascó even provided a Web site where the curious could compare search results between Google Scholar and the likes of Nature, Wiley, or Blackwell.[6] More recently, studies have appeared that track the "value-added" open-access citations that appear uniquely in Google Scholar versus other sources.[7] However, is comprehensiveness of content the primary indicator of a resource's usefulness?

Every title from every database may not be in Google Scholar, but that should not be an indictment of Google Scholar's inability to return scholarly results across disciplines. The algorithms Google Scholar uses to return result sets cannot really be compared to library database algorithms. However, what is returned can be judged for its relevancy and scholarliness. Up to this point, studies of Google Scholar have followed the example of Neuhaus et al., which compared Google Scholar content to forty-seven other databases.[8] This title-by-title and citation-by-citation comparison is a pure numerical measure, but it neglects to address the efficacy of any particular search or the scholarly nature of content or algorithms in discovering that content.

We felt that a different approach was needed. Rather than measuring what has gone into the database, we have sought, to some degree, to evaluate what comes out as a result of search queries. We have done this by involving subject librarians with knowledge of typical reference questions and using those questions to query both Google Scholar and discipline-specific databases. We then asked the same librarians to judge the search results using a rubric of scholarliness. In short, we wanted to determine how appropriate it would

be to include each citation in a scholarly research paper at an academic institution.

This notion of scholarliness, that we attempted to encapsulate in the rubric, uses a common collection-assessment tool as outlined by Kapoun.[9] This model considers many factors, including accuracy, authority, objectivity, currency, and coverage. For the purposes of the study, we added relevancy because materials that met those five criteria were not always relevant to the research topic.

The Kapoun model of evaluating scholarliness was based on his experience in evaluating print resources but was expanded for evaluating Web resources. Because this study was constructed to compare library database results to Google Scholar results, we had no way of knowing the breadth of materials the rubric would be required to evaluate. Google Scholar alone references materials in any format whether it is in print or electronic only and includes journals, books, syllabi, and conference proceedings. These are just a few examples of the disparate types of materials the rubric would need to handle. By using a model flexible enough to evaluate materials in any format, we have attempted to inject a qualitative value of Google Scholar results to the ongoing debate.

Methodology

We selected seven subject librarians from Brigham Young University to cover various academic disciplines: humanities, sciences, and social sciences. Each specialist was blind to the purpose of the study. We requested that they provide us (1) a sample question that they typically receive from students, (2) a structured query to search a library database, and (3) the library database they would use for that particular query.

We then used their data in two different ways. First, we translated the library database query into an equivalent search string used by Google Scholar. Using the original query and the query translated to work with Google Scholar, we searched

	TABLE 1		
	Academic Representation in This Study		
Academic Discipline	**Database Query**	**GS Query**	**Library Database**
Science	(ACL or "anterior cruciate ligament*") and injur* and (athlet* or sport or sports) and (therap* or treat* or rehab*)	ACL OR "anterior cruciate ~ligament" ~injury ~athlete OR sport ~therapy OR ~treatment OR ~rehabilitation	SportDiscus
Science	lung cancer and (etiol* or caus*) and (cigarette* or smok* or nicotine*)	lung cancer ~etiology OR ~cause ~cigarette OR ~smoking OR ~nicotine	Medline
Science	"dark matter" and evidence	"dark matter" evidence	Applied Science and Technology Abstracts
Social Science	("fast food" or mcdonald's or wendy's or "burger king" or restaurant) and franchis* and (knowledge n3 transfer or "knowledge management" or train*)	"fast food" OR mcdonald's OR wendy's OR "burger king" OR restaurant ~franchise "knowledge transfer" OR "knowledge management" OR ~train	Business Source Premier
Social Science	("standardized test*" or "high stakes test*") and ("learning disabilit*" or dyslexia or "learning problem") and accommodat*	"standardized ~test" OR "high stakes ~test" "learning ~disability" OR dyslexia OR "learning problem" ~accommodation	PsycINFO
Humanities	(bilingual* or L2) and (child* or toddler) and "cognitive development"	~bilingual OR L2 ~child OR toddler "cognitive development"	Linguistics and Language Behavior Abstracts
Humanities	(memor* or remembrance or memoir*) and (holocaust) and (Spiegelman or Maus)	~memor OR remembrance OR ~memoir holocaust Spiegelman OR Maus	JSTOR

both the library database and Google Scholar and retrieved the citations and full text for the first thirty results. We selected thirty results because research has shown that less than one percent of all users ever go beyond a third page of results and most search engines return about ten results per page.[10]

Next, we took the citations from the library databases and determined if they could also be found using Google Scholar and took the citations from Google Scholar to see if they could also be found

in the library database. This allowed us to calculate the overlap of citations between the library databases and Google Scholar.

We standardized the formatting of the citations and inserted them randomly into a spreadsheet, which contained a rubric that was used to assign a scholarliness score to each of the citations. The rubric contained six criteria, based on Kapoun's model of evaluating resources, to judge scholarliness: (1) accuracy, (2) authority, (3) objectivity, (4) currency, (5) coverage, and (6) relevancy.[11] These criteria were

TABLE 2 Rubric for Grading Scholarliness							
1 = Below Average Quality; 2 = Average Quality; 3 = Above Average Quality							
Citation Number	References	Accuracy	Authority	Objectivity	Currency	Coverage	Relevancy
1	Barnes, J.E., & Hernquist, L.E. (1993) Computer models of col-liding galaxies. Physics Today, 46, 54–61.	1 2 3	1 2 3	1 2 3	1 2 3	1 2 3	1 2 3
2	Bergstrom, L. (2000) Nonbary-onic dark matter: Observational evidence and de-tection methods. Reports on Prog-ress in Physics, 63(5), 793–841.	1 2 3	1 2 3	1 2 3	1 2 3	1 2 3	1 2 3

graded on a scale of 1 (below average) to 3 (above average) and summed to create a total scholarliness score for each citation.

We provided the subject librarians with the full text of each of the citations and asked them to use the rubric to evaluate the scholarliness of the individual citations. After the grading was completed, we were able to group each citation from the subject librarian into one of three categories: (1) the citation was available only in the library database, (2) the citation was available only in Google Scholar, or (3) the citation was available in both the library database and Google Scholar. We have used the term "exclusivity" to describe the three categories.

Once we had grouped the citations by category, we ran a statistical analysis that controlled for the effect of the individual librarian on the total scholarliness score, for the effect of "exclusivity" and for any interaction there may have been between both librarian and "exclusivity":

$$total\ scholarliness\ score = \mu + E_i + L_j + EL_{ij} + \varepsilon_{ijk}$$
Where:
μ = Average total score

E = Effect due to "exclusivity" (i = 1, 2, 3)
L = Effect due to librarian (j = 1, 2, …7)
EL = Interaction between "exclusivity" and librarian
ε = Error term (k = degrees of freedom associated with the error term)

Within the context of this formula, E controls for any effect due to the "exclusivity" of the citation, where i represents each of the three categories of "exclusivity" (that is, the citation was found only in the database, it was found only in Google Scholar, or it was found in both the database and Google Scholar). Each librarian (L) also played a role in the total scholarliness score. One librarian could have provided consistently low scores, with another having a tendency toward higher scores. To account for this disparity, each librarian was treated as a factor in the total scholarliness score, where j represents each of the seven participants. In short, this formula allowed us to calculate a measure of scholarliness while accounting for differences in the location of citations as well as differences between librarians.

TABLE 3
Scholarliness Based on "Exclusivity" (Maximum Scholarliness Score Is 18)

Participant	Found Only in Database Average Score	Found Only in GS Average Score	Percent Change in Scholarliness Score Between the Database and GS	Found in Both Average Score
1	11.7	16.1	36.8%	13.5
2	13.2	13.8	4.5%	14.6
3	N/A	12.0	N/A	15.6
4	10.0	13.5	35.0%	14.3
5	10.0	11.6	16.0%	11.5
6	11.7	12.8	8.5%	14.3
7	16.5	14.4	−12.7%	13.9
Least Squares Mean	11.9	14.0	17.6%	14.2

Results

The mean scholarliness score of citations found only in Google Scholar was 17.6 percent higher than the score for citations found only in licensed library databases. In fact, across all but one of the tested disciplines, citations found only in Google Scholar had a higher average scholarliness score than citations found only in licensed library databases. The one discipline with a lower score, however, had only two unique citations in the library database, so the exact significance of the scores for that discipline is imprecise. Additionally, the citations found in both Google Scholar and licensed library databases had a higher average score than citations found only in one or the other. Finally, there was no statistically significant difference found between the scholarliness score across disciplines within Google Scholar. Searching for either a humanities topic or a science topic yielded no difference in the scholarliness score of citations discovered in Google Scholar.

Discussion of Results

It is interesting to note that there was very little overlap between the initial thirty citations returned by the databases and the initial thirty citations returned by Google Scholar. In fact, only one query of the seven had any overlapping citations between Google Scholar and the database—an overlap of five citations from JSTOR that appeared within the first thirty results in Google Scholar.

However, during the second phase of the study, when we began to search for specific citations, we found that Google Scholar actually contained 76 percent of all the citations found in the library databases, while the library databases contained only 47 percent of the citations found in Google Scholar. Despite the initial lack of overlap in the search results, it was clear that Google Scholar included a large portion of the citations available in library databases.

TABLE 4
Overlap of Citations

Participant	Percent of Database Citations in GS	Percent of GS Citations in Database
1	76.7%	0.0%
2	83.3%	43.3%
3	100.0%	96.7%
4	96.7%	80.0%
5	93.3%	28.0%
6	0.0%	46.7%
7	81.8%	34.5%
Average	76.0%	47.0%

This seems to validate the decision of many students to use Google first to look for information. If Google Scholar contains much of the content available in library databases, why shouldn't students begin where the most content exists? The argument is made that Google Scholar will return millions of hits, many of which are spurious at best, while a library database will only return a few thousand results that are more focused to the query.

However, the power of ordering results by relevancy, combined with the fact that very few people ever go beyond the third page of results, creates a searcher-imposed higher level of precision for any search engine. This is particularly true of Google Scholar, where the most relevant and more scholarly material floats to the top of the list, while the less precise material falls to the bottom, where it is rarely seen. Hit counts are of secondary importance in a Google Scholar search; the key to Google Scholar's success is relevancy ranking and a large universe of information.

A database is limited to its defined title list of content, whereas Google Scholar, by its very nature, is open to a much broader set of content that aids the researcher. Business Source Premier, one of the library databases, was the only library database where we found more Google Scholar citations in the database than database citations in Google Scholar. However, even in this one instance, the scholarly score for citations found only in Google Scholar was higher than the score for citations found only in the database. The citations found in both Google Scholar and the database received even higher scores, and these citations were only exposed through the first thirty hits in Google Scholar. This seems to indicate that, even when Google Scholar is returning fewer titles, as in this case with Business Source Premier, it still returns citations that are more scholarly to the top.

Up to this point, many library databases have defaulted to sorting by date rather than by relevancy. The fact that many databases are now adding rel-

evancy search options seems to indicate that Google Scholar got it right in the first place. It appears that Google Scholar has done a better job of both precision and recall than library databases have.

Many studies have compared content in library databases to content in Google Scholar and found inconsistencies. The purpose of both search systems, however, is to discover relevant, scholarly content. Using our scholarliness model, we found that, across disciplines, Google Scholar is generally superior to individual databases in retrieving appropriate citations. As more publishers share their content with Google Scholar, we would expect the effectiveness of a Google Scholar search to increase.

Future Studies

The statistical results from this study can be extrapolated only to the specific topics and subject librarians that were involved in the study. A more comprehensive statistical methodology would need to be constructed to make the results generally applicable. However, our results were compelling enough to make us believe that the results would hold up to more strenuous tests. Additionally, the rubric we used in our study was only a three-point Likert scale. Finding statistically significant differences would have been easier had we selected a seven or more point Likert scale.

Additionally, our analysis used a vetted approach to evaluating scholarliness of resources. A more objective view of scholarliness could be obtained by using some variation of citation analysis (such as citation counts or ISI impact factor). We started to do such an analysis but decided there were too many trade-offs to be appropriate, given the methodology we used for this study. For example, citation counts are difficult to come by for materials other than journal articles, and impact factors are calculated for journals only and not for specific articles.[12] Alternate methodologies might be able to overcome or account for the shortcomings of using citation analysis to judge scholarliness.

Our study used skilled librarians to create search queries and to judge the quality of the citations retrieved. Unlike most students, the librarians used complex search queries to find more relevant results. Students would be more likely to use natural language queries to find citations. Complex search queries could return very different results from natural language queries. Future studies will need to address the potential differences to find out if the results we found hold across different types of searches.

Finally, future studies need to look at the appropriateness of comparing Google Scholar to individual library databases. It is probable that federated searching is more comparable to Google Scholar than are individual library databases. However, how users and librarians select which resources to use in a federated search and how the federated search engine returns the results would still impact the discoverability of scholarly resources. Some studies have already started down this road,[13] but Google Scholar result sets have still not been carefully compared to result sets from federated search products.

Libraries have begun to build local Google Scholars, using tools such as *Primo* (Ex Libris), *AquaBrowser* (Medialab Solutions), and *Encore* (Innovative Interfaces), that have the potential to aid users in discovering even more scholarly materials than what is currently discovered in Google Scholar. Comparing Google Scholar to a future system that has completely indexed all local content and content available to libraries but provided by third parties would be the ultimate comparison.

Conclusion

Typical arguments against Google Scholar focus on citation counts and point to inconsistent coverage between disciplines. We felt the more appropriate analysis was to compare the scholarliness of resources discovered using Google Scholar with resources found in library databases. This analysis showed that Google Scholar yielded more scholarly content than library databases, with no statistically significant difference in scholarliness across disciplines. Despite these findings, Google Scholar is not in competition with library databases. In truth, without the cooperation of database vendors and publishers, Google Scholar would not exist as it does today. Google Scholar is simply a discovery tool for finding scholarly information, while databases still perform the function of providing access to the content unearthed by a Google Scholar search. The enhanced discoverability of information in Google Scholar makes it a great tool for librarians as well as library users.

Notes

1. Jan Brophy and David Bawden, "Is Google Enough? Comparison of an Internet Search Engine with Academic Library Resources," *Aslib Proceedings* 57, no. 6 (2005): 498–512; Susan Gardner and Susanna Eng, "Gaga Over Google? Scholar in the Social Sciences," *Library Hi Tech News* 22, no. 8 (2005): 42–45; Martin Kesselman and Sarah Barbara Watstein, "Google Scholar and Libraries: Point/Counterpoint," *Reference Services Review* 33, no. 4 (2005): 380–87.

2. Nisa Bakkalbasi et al., "Three Options for Citation Tracking: Google Scholar, Scopus and Web of Science," *Biomedical Digital Libraries* 3, no. 7 (2006), available online at www.bio-diglib.com/content/3/1/7 [accessed 24 March 2008]; Kayvan Kousha and Mike Thelwall, "Google Scholar and Google Web/URL Citations: A Multi-Discipline Exploratory Analysis," *Journal of the American Society for Information Science and Technology* 58, no. 7 (2007): 1055–65; Mary L. Robinson and Judith Wusteman, "Putting Google Scholar to the Test: A Preliminary Study," *Program: Electronic Library & Information Systems* 41, no. 1 (2007): 71–80; Chris Neuhaus et al., "The Depth and Breadth of Google Scholar: An Empirical Study," *Portal: Libraries and the Academy* 6, no. 2 (2006): 127–41.

3. Jeffrey Pomerantz, "Google Scholar and 100 Percent Availability of Information," *Information Technology and Libraries* 25 (June 2006): 52–56.

4. Laura Bowering Mullen and Karen A. Hartman, "Google Scholar and the Library Web Site: The Early Response by ARL Libraries," *College and Research Libraries* 67, no. 2 (2006): 106–22.

5. Péter Jascó, "Deflated, Inflated and Phantom Citation Counts," *Online Information Review*

30, no. 3 (2006): 297–309; Péter Jascó, "As We May Search: Comparison of Major Features of the Web of Science, Scopus, and Google Scholar Citation-Based and Citation-Enhanced Databases," *Current Science* 89, no. 9 (2005): 1537–47; Péter Jascó, "Google Scholar: The Pros and the Cons," *Online Information Review* 29, no. 2 (2005): 208–14.

    6. Péter Jascó, "Side-by-side2: Native Search Engines vs Google Scholar" (2005). Available online at www2.hawaii.edu/~jacso/scholarly/side-by-side2.htm. [Accessed 24 March 2008].

    7. Kayvan Kousha and Mike Thelwall, "How Is Science Cited on the Web? A Classification of Google Unique Web Citations," *Journal of the American Society for Information Science and Technology* 58, no. 11 (2007): 1631–44.

    8. Neuhaus et al., "Depth and Breadth of Google Scholar."

    9. Jim Kapoun, "Teaching Undergrads Web Evaluation: A Guide for Library Instruction," *College and Research Libraries News* 59, no. 2 (1998): 522–23.

    10. Jakob Nielsen and Hoa Loranger, *Prioritizing Web Usability* (Berkeley, Calif.: New Riders, 2006).

    11. Kapoun, "Teaching Undergrads Web Evaluation."

    12. Per O. Seglen, "Why the Impact Factor of Journals Should Not Be Used for Evaluating Research," *BMJ* 314, no. 7079 (1997): 498–502.

    13. Xiaotian Chen, "Metalib, Webfeat, and Google: The Strengths and Weaknesses of Federated Search Engines Compared With Google," *Online Information Review* 30, no. 4 (2006): 413–27.