Charleston Library Conference

Summon, EBSCO Discovery Service, and Google Scholar: Comparing Search Performance Using User Queries

John Vickery
North Carolina State University Libraries

Karen Ciccone
Science Informatics

John Vickery and Karen Ciccone, "Summon, EBSCO Discovery Service, and Google Scholar: Comparing Search Performance Using User Queries" (2015). *Proceedings of the Charleston Library Conference.*
http://dx.doi.org/

Summon, EBSCO Discovery Service, and Google Scholar: Comparing Search Performance Using User Queries

John Vickery, Analytics Coordinator & Collection Manager for Social Sciences, NCSU Libraries

Karen Ciccone, Director Natural Resources Library & Research Librarian, Science Informatics

Abstract

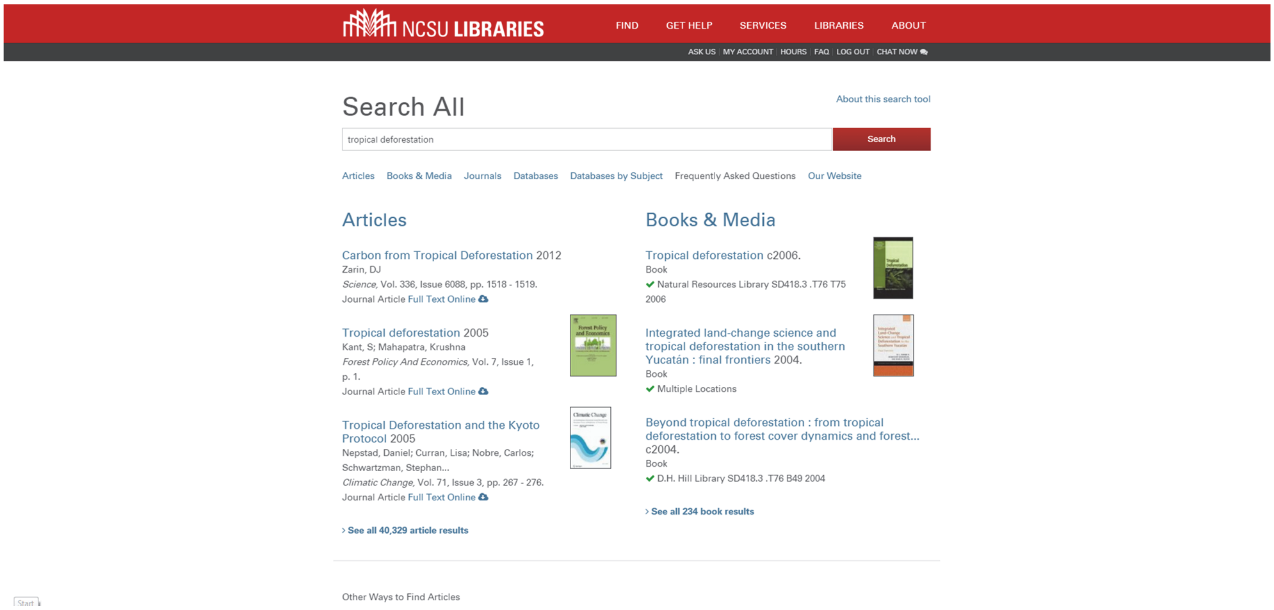
When the NCSU Libraries initially subscribed to the Summon Discovery Service in 2009, there were few other competitors on the market and none offered an API interface that could be used to populate the “Articles” portion of our QuickSearch application (http://search.lib.ncsu.edu/). Since then, EBSCO Discovery Service (EDS) has emerged as a viable competitor. Using a random sample of actual user searches and bootstrap randomization tests (also referred to as permutation tests), the NCSU Libraries’s Web-Scale Discovery Product Team conducted a study to compare the search performance of Summon, EDS, and Google Scholar.

Introduction

This paper discusses a study done by the NCSU Libraries Web-Scale Discovery Product Team to compare the search performance of Summon, EBSCO Discovery Service, and Google Scholar. The paper provides background on the study and its objectives. We then discuss the study methods and results. This paper and the accompanying presentation at the 2015 Charleston Conference are drawn from a detailed article about the project published by Karen Ciccone and John Vickery in *Evidence Based Library and Information Practice, 10*(1) (2015). Readers are encouraged to consult that article for a more detailed discussion of the study. In addition, the SAS code written for the analysis is discussed in more detail in a paper by John Vickery in the 2015 SAS Global Forum Proceedings (2015). The code is also made available on GitHub at https://github.com /jnvickery/permutations.

Project Background

Similar to many libraries, the NCSU Libraries has invested substantial time, effort, and money into



Articles

Carbon from Tropical Deforestation 2012
Zarin, DJ
Science, Vol. 336, Issue 6088, pp. 1518 - 1519.
Journal Article Full Text Online

Tropical deforestation 2005
Kant, S; Mahapatra, Krushna
Forest Policy And Economics, Vol. 7, Issue 1, p. 1.
Journal Article Full Text Online

Tropical Deforestation and the Kyoto Protocol 2005
Nepstad, Daniel; Curran, Lisa; Nobre, Carlos; Schwartzman, Stephan...
Climatic Change, Vol. 71, Issue 3, pp. 267 - 276.
Journal Article Full Text Online

› See all 40,329 article results

Books & Media

Tropical deforestation c2006.
Book
✓ Natural Resources Library SD418.3 .T76 T75 2006

Integrated land-change science and tropical deforestation in the southern Yucatán : final frontiers 2004.
Book
✓ Multiple Locations

Beyond tropical deforestation : from tropical deforestation to forest cover dynamics and forest... c2004.
Book
✓ D.H. Hill Library SD418.3 .T76 B49 2004

› See all 234 book results

Other Ways to Find Articles

Figure 1. The NCSU Libraries’s QuickSearch interface.

http://dx.doi.org/10.5703/1288284316288

implementing a web-scale discovery service. Given this investment, it is important to periodically review competing products to see if they would be a better fit for us either because of price or effectiveness. Our Web-Scale Discovery Product Team is charged with testing and evaluating our existing Summon service as well as other potential products. The team is made up of nine librarians representing public services, technical services, collections, and IT.

At NCSU, we primarily use the Summon API to populate the “Articles” section of our QuickSearch as shown in Figure 1. A search in the QuickSearch application presents separate results for Articles, Books & Media, Our Website, and other information.

In 2009, when the NCSU Libraries implemented Summon, it was the only product that had an API that could be leveraged to be used with our QuickSearch application. Since 2009, other products including EBSCO Discovery Service (EDS) have implemented API functionality. As such, the Web-Scale Discovery Product Team decided a comparison was warranted and a trial for EDS was set up in April and May of 2014 in order to conduct the study.

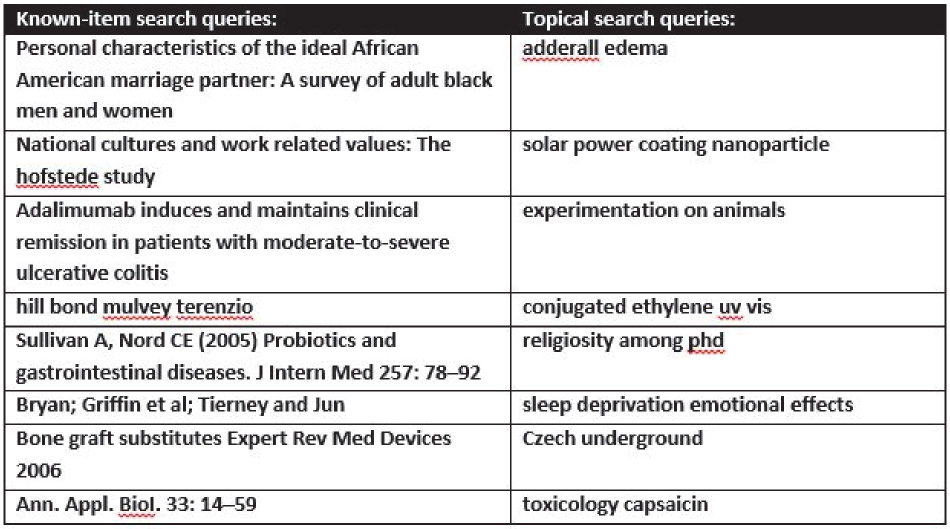
Other studies have, of course, been conducted to assess the performance of discovery services. In 2013, Asher, Duke, and Wilson compared Summon, EBSCO Discovery Service, and Google Scholar on the basis of search performance. Scores were based on librarian quality ratings of the resources selected by test subjects. Quality of resources was based on whether articles were from scholarly or non-peer-reviewed journals.

In 2013, Rochkind compared user preference for search results from the major discovery systems including EDS, Summon, EBSCOhost “Traditional” API, Ex Libris Primo, and Elsevier’s Scopus. Users were presented with side-by-side views of results from two randomly chosen products. Users were then asked to indicate which set of results they preferred, with an option for “Can’t Decide/About the Same.”

Both of the above studies relied on test subjects entering hypothetical searches. The NCSU study differs in that actual user queries from the Summon log files were used. This provides a sample of actual research questions as well as a distribution of searches that accurately reflects the relative frequency of known-item to topical searches.

The primary objective of the study was to answer the question of whether Summon or EDS produced better results for the type of searches typically performed by our users. We can get a sense of what is “typical” by examining our Summon search logs. These logs contain the actual queries that users entered into the search box. From looking at these logs, we know that about three quarters of the searches are for “topical” type searches with the rest being “known-item” type searches where a user pastes in a citation or clearly enters title and author keywords. Because of how these two types of searches differ, the study was designed to separately evaluate how the discovery systems handled these types of

Table 1. Examples of known-item and topical search queries.



Known-item search queries:	Topical search queries:
Personal characteristics of the ideal African American marriage partner: A survey of adult black men and women	adderall edema
National cultures and work related values: The hofstede study	solar power coating nanoparticle
Adalimumab induces and maintains clinical remission in patients with moderate-to-severe ulcerative colitis	experimentation on animals
hill bond mulvey terenzio	conjugated ethylene uv vis
Sullivan A, Nord CE (2005) Probiotics and gastrointestinal diseases. J Intern Med 257: 78–92	religiosity among phd
Bryan; Griffin et al; Tierney and Jun	sleep deprivation emotional effects
Bone graft substitutes Expert Rev Med Devices 2006	Czech underground
Ann. Appl. Biol. 33: 14–59	toxicology capsaicin

searches. Table 1 lists examples of known-item and topical search queries from the Summon logs.

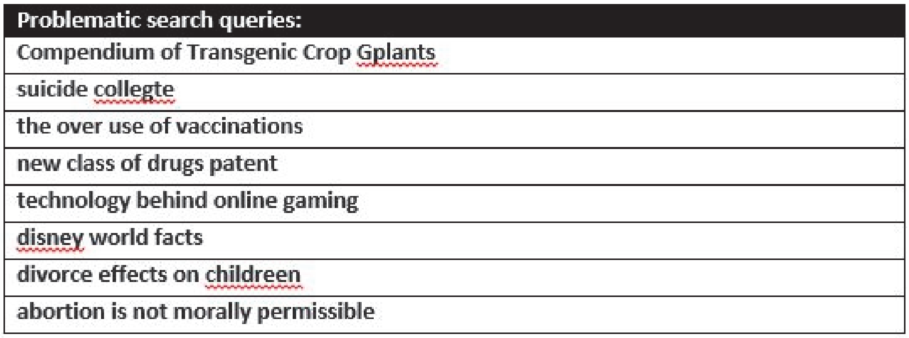
Discovery services such as Summon and EDS can offer several advantages over Google Scholar (e.g., API available, ability to save and e-mail results, ability to limit to peer-reviewed articles). However, we know that the Google universe is often the first choice of our users. Therefore, we were interested in how Google Scholar would perform as compared to Summon and EDS.

Methods

The study used actual user search queries from our Summon search logs. This differentiates the NCSU study from both the Asher and Rochkind studies (2013; 2013). The user queries occasionally contain typos, punctuation errors or extraneous characters. In addition, the queries could also be overly broad or otherwise problematic. Table 2 lists examples of problematic search queries from our sample.

The SAS software PROC SURVEYSELECT procedure was used to generate a simple random sample of 225 search queries. The sample was drawn from the approximately 664,000 Summon searches performed between January 1, 2013 and December 31, 2013. We determined the sample to be large enough to be representative of the population but also small enough to be manageable by the Web-Scale Discovery Product Team. Each of the nine members of the team was given twenty-five queries to analyze. Two team members were unable to complete testing, resulting in 183 queries tested.

Table 2. Examples of problematic search queries from our sample.



Problematic search queries:
Compendium of Transgenic Crop Gplants
suicide collegte
the over use of vaccinations
new class of drugs patent
technology behind online gaming
disney world facts
divorce effects on childreen
abortion is not morally permissible

Team members coded queries as topical search queries or known-item search queries. Team members entered each query into Summon, EDS, and Google Scholar. For topical searches, team members recorded the number of relevant results within the first ten results. A result was considered relevant if it matched the presumed topic of the user's search. Only the title and abstract were used to determine relevancy.

For known-item search queries, team members coded "yes" or "no" responses to the questions of "Did you find the item?" and "Was it in the top three results?"

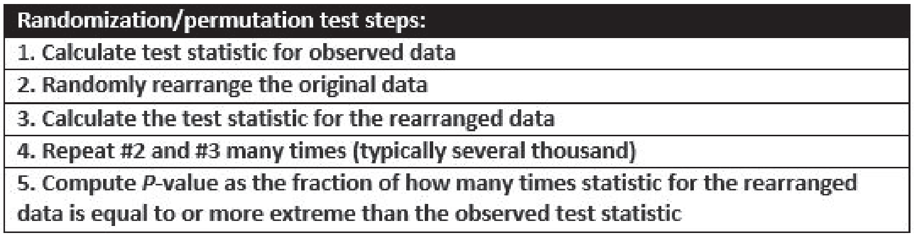
The topical search queries were analyzed graphically as well as with a permutation test for repeated measures analysis of variance and pairwise permutation tests for comparing the means of the paired data. The permutation test code was written in SAS/IML software. The known-item search queries were analyzed using a graphical comparison of the number of found known items and a Mantel-Haenszel analysis to examine the relationship between discovery product and success of a known-item search.

About Permutation Tests

Permutation tests are defined in *A Dictionary of Statistics* (Upton & Cook, 2014) as follows:

> A simple type of hypothesis test. Denote the value of some test statistic by *T*. The observed data values are randomly redistributed amongst the experimental units. The test statistic is calculated for each such redistribution. Depending on the number of data values, either all possible permutations are made, or a random selection (of say 1,000 permutations) is made. For each permutation the value of the test statistic is considered. The significance of the value *T* is determined by the proportion of permutations that lead to values greater than, or equal to, *T*.

Table 3. Basic steps for a randomization/permutation test.



Randomization/permutation test steps:
1. Calculate test statistic for observed data
2. Randomly rearrange the original data
3. Calculate the test statistic for the rearranged data
4. Repeat #2 and #3 many times (typically several thousand)
5. Compute *P*-value as the fraction of how many times statistic for the rearranged data is equal to or more extreme than the observed test statistic

The concept of permutation tests was discussed as early as the 1930s by Fisher and others (Anderson, 2001; Good, 2005). However, the practicality of the method was limited by computing power. As the definition above states, the basic premise of a permutation test is that by randomly shuffling the data and recalculating a test statistic, a permutation test can calculate the probability of getting a value equal to or more extreme than an observed test statistic. Edgington and Onghena provide a detailed overview of using randomization tests to determine *P*-values in repeated measures designs (2007).

Results

For the comparison of known-item searches, the proportion of items found using Summon, EDS, and Google Scholar was nearly the same. Figure 2 shows the frequency of known items found between the three products.

For the question of whether or not a found item was in the top three results, the products performed nearly the same. All but two of the found known items were in the top three results for EDS and Google Scholar, and all but one was in the top three for Summon.

Adjusting or controlling for the sample query, no significant difference was found between Summon, EDS, and Google Scholar success rates, χ2 (2, *N* = 132) = 0.08, *p* = 0.96. The small sample size for known items (*n* = 44) was a limitation of our study. The test did not have the sufficient power to detect small differences (< 40%) in the performance of the products. We would suggest that subsequent studies use a larger starting sample in order to obtain sufficient known-item queries to detect small performance differences between products.

For the comparison of topical searches, the three products also performed similarly. Figure 3 shows the distribution of the number of relevant results for Summon, EDS, and Google Scholar.

Table 4 lists the mean number of relevant results for each product. The mean for Summon and EDS is nearly the same. Google Scholar appears to have on average approximately one additional relevant result.

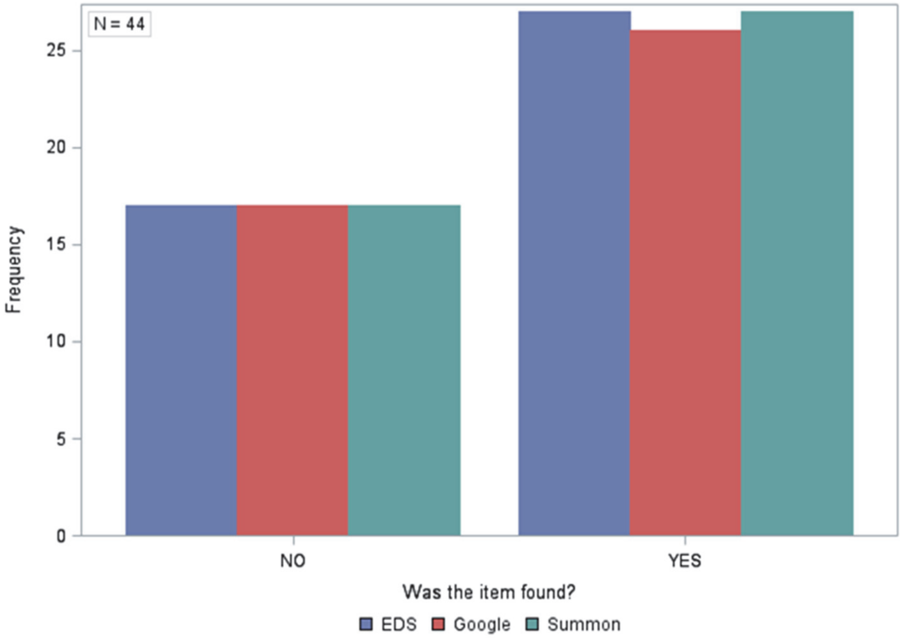


Figure 2. Frequency of known items found for EDS, Google Scholar, and Summon.

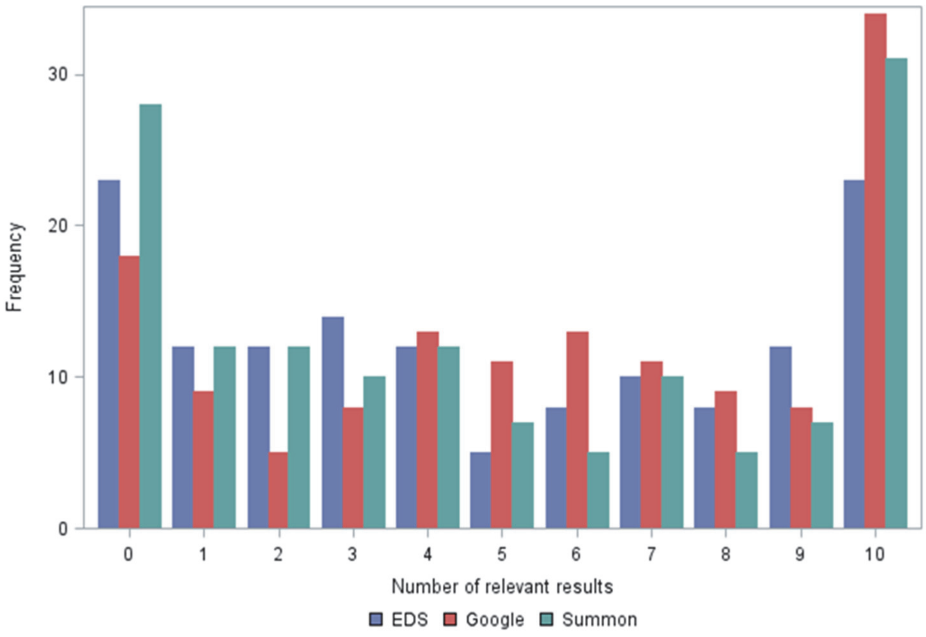


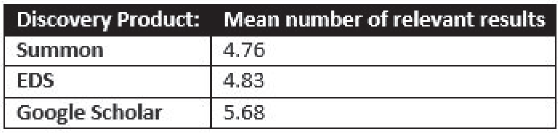
Figure 3. Frequency distribution of the number of relevant results from EDS, Google Scholar, and Summon.

In order to detect any overall difference between the means, we performed a permutation test for repeated measures analysis (Good, 2005; Howell, 2006). Figure 4 (see Appendix) shows the results of ten thousand simulations of the *F*-statistic. The small *P*-value of 0.002 indicates that there is an overall difference in the means.

Given the indication of an overall difference in the mean number of results between the three products, we performed pairwise permutation tests to confirm where the difference occurred. The tests compared each of the possible pairs: Summon to EDS, EDS to Google Scholar, and Summon to Google Scholar.

There was no significant difference in the mean number of relevant results between Summon and EDS. Figure 5 (see Appendix) shows the permutation test results comparing Summon and EDS. The large *P*-value of 0.773 represented by the red shading of the histogram indicates no significant difference between the two products.

Table 4. Mean number of relevant results for each discovery product.



Discovery Product:	Mean number of relevant results
Summon	4.76
EDS	4.83
Google Scholar	5.68

There was, however, a significant difference between the mean number of relevant results for Google Scholar and both EDS and Summon. In our observed data, Google Scholar outperformed EDS by an average of 0.85 relevant results. Figure 6 (see Appendix) shows a histogram of the results of the permutation test. As indicated by the small *P*-value of 0.004, the test indicates that it is highly unlikely that this difference was due to chance alone.

Similarly, Google Scholar outperformed Summon by an average of 0.91 relevant results. Figure 7 (see Appendix) shows the results of the comparison between Summon and Google Scholar. Here again, the *P*-value associated with the test is 0.004 and indicates a significant difference in the mean number of relevant results between Summon and Google Scholar.

Conclusion

The NCSU Libraries's Web-Scale Discovery Product Team conducted a study to compare the search performance of Summon, EBSCO Discovery Service (EDS), and Google Scholar. A random sample of 183 actual user searches from the NCSU Libraries's 2013 Summon search logs was used for the study. No significant difference in performance between Summon and EDS for either known-item or topical searches was found. There was also no significant difference between the Summon, EDS, and Google Scholar for known-item searches. However, Google Scholar outperformed both Summon and EDS for topical searches. Given the lack of significant difference in performance between Summon and EDS, NCSU Libraries's decision to purchase one product or the other can be based upon other considerations such as technical issues, cost, customer service, or user interface.

References

Anderson, Marti J. (2001). "Permutation Tests for Linear Models." *Australian & New Zealand Journal of Statistics, 43*(1), 75–88.

Asher, A. D., Duke, L. M., & Wilson, S. (2013). Paths of discovery: Comparing the search effectiveness of Ebsco Discovery Service, Summon, Google Scholar, and conventional library resources. *College & Research Libraries, 75*(5), 464–488.

Ciccone, K., & Vickery, J. (2015). Summon, EBSCO Discovery Service, and Google Scholar: A comparison of search performance using user queries. *Evidence Based Library And Information Practice, 10*(1), 34–49. Retrieved from http://ejournals.library.ualberta.ca/index.php/EBLIP/article/view/23845

Edgington, E. S., & Onghena, P. (2007). *Randomization tests*. Boca Raton: Chapman & Hall/CRC.

Good, P. (2005). *Permutation, parametric and bootstrap tests of hypotheses* (3rd ed.). New York: Springer.

Howell, D. C. (2006). *Repeated measures analysis of variance via randomization*. Retrieved from https://www.uvm.edu/~dhowell/StatPages/Resampling/RandomRepeatMeas/RepeatedMeasuresAnova.html

Rochkind, J. (2013). A comparison of article search APIs via blinded experiment and developer review. *Code4Lib Journal, 19*. Retrieved from http://journal.code4lib.org/articles/7738

Upton, G., & Cook, I. (2014). Permutation test. In *A Dictionary of Statistics*. New York: Oxford University Press. Retrieved from http://www.oxfordreference.com/view/10.1093/acref/9780199679188.001.0001/acref-9780199679188-e-2108

Vickery, J. (2015). Permit me to permute: A basic introduction to permutation tests with SAS/IML. *SAS Global Forum 2015 Proceedings*. Retrieved from http://support.sas.com/resources/papers/proceedings15/2440-2015.pdf

Appendix

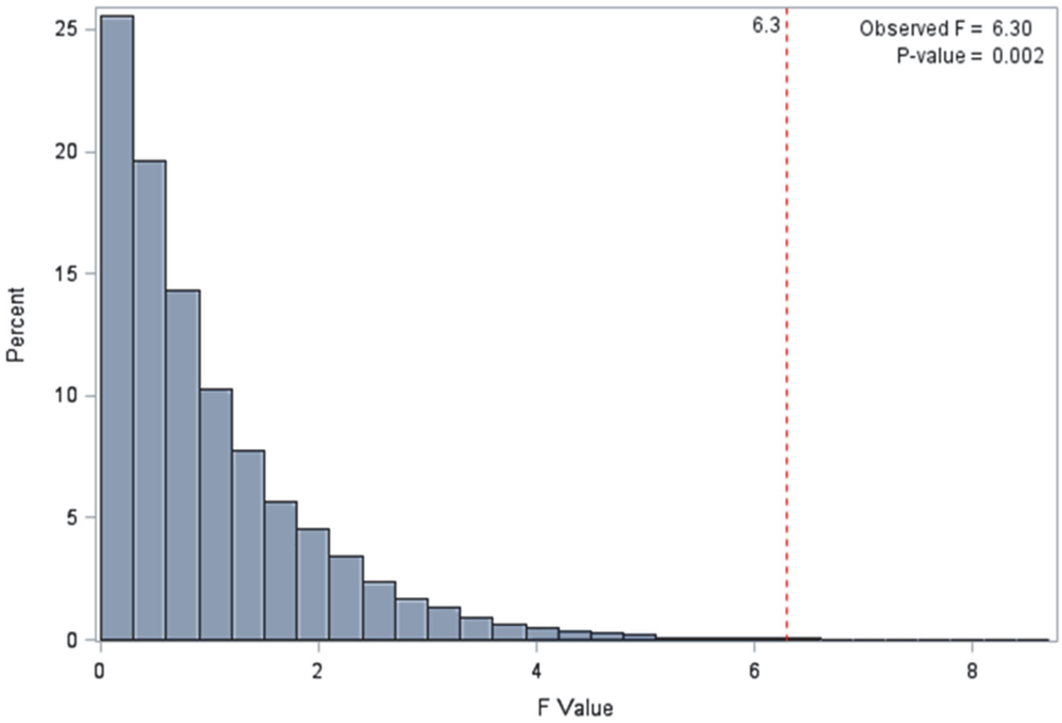


Figure 4. Bootstrap distribution of the *F*-statistic under the null hypothesis for 10,000 resamples indicates an overall difference between the mean numbers of relevant results for EDS, Google Scholar, and Summon.

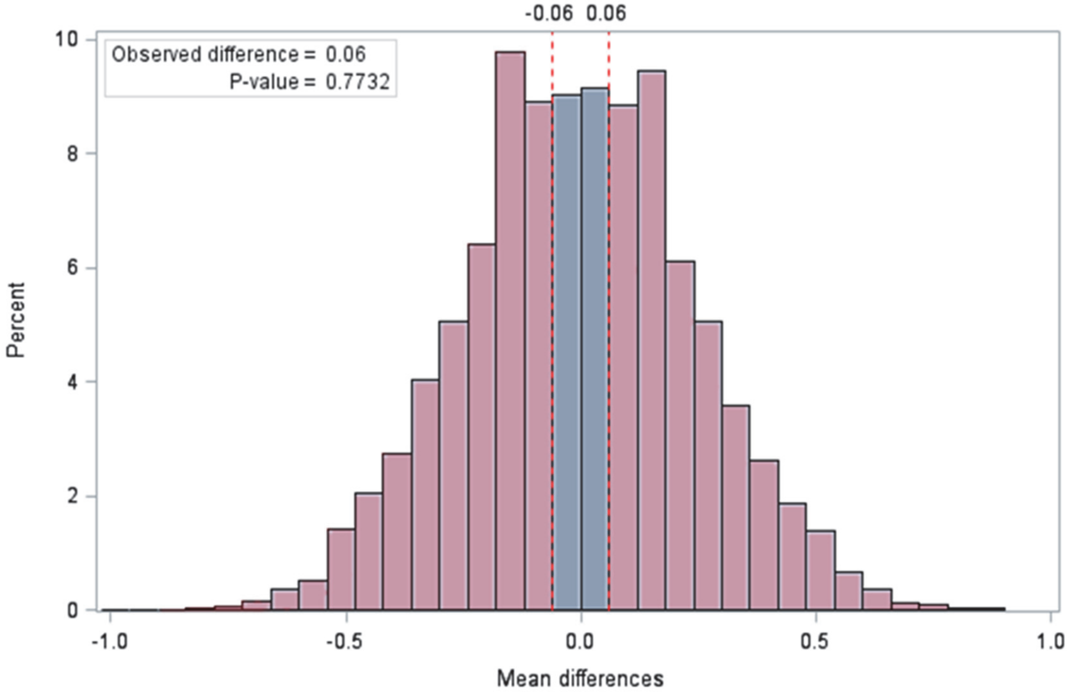


Figure 5. Bootstrap distribution under null hypothesis for 10,000 resamples shows that the observed difference in the mean number of relevant results between Summon and EDS was not significant.

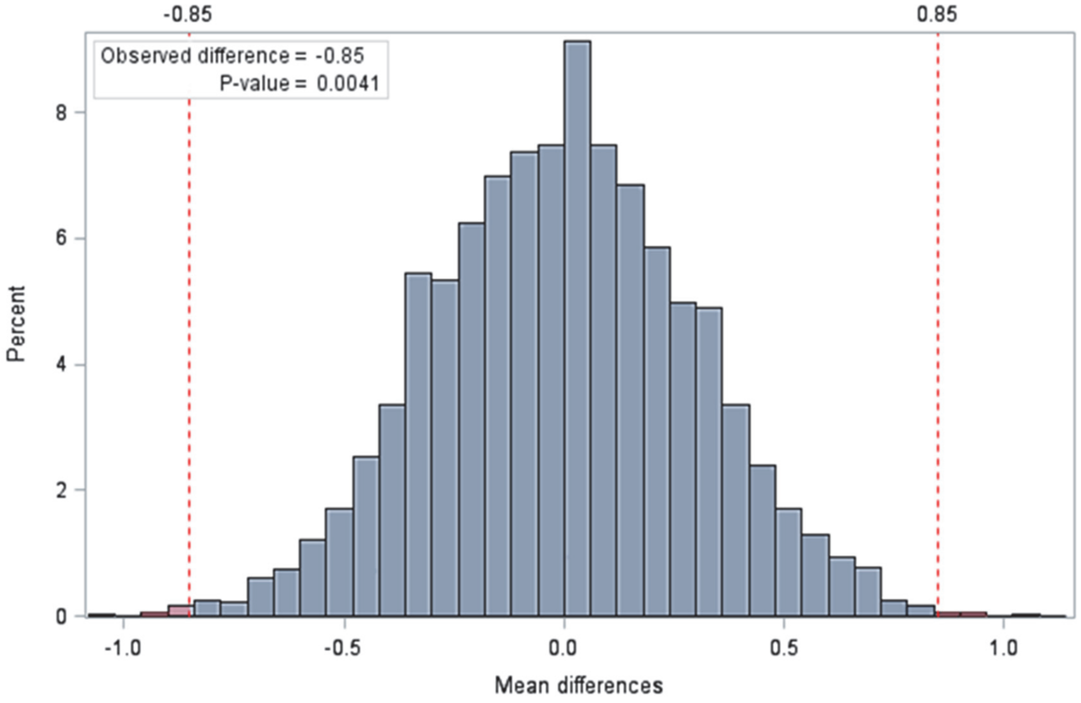


Figure 6. Bootstrap distribution under the null hypothesis for 10,000 resamples shows that the observed difference in the mean number of relevant results between Google Scholar and EDS was significant.

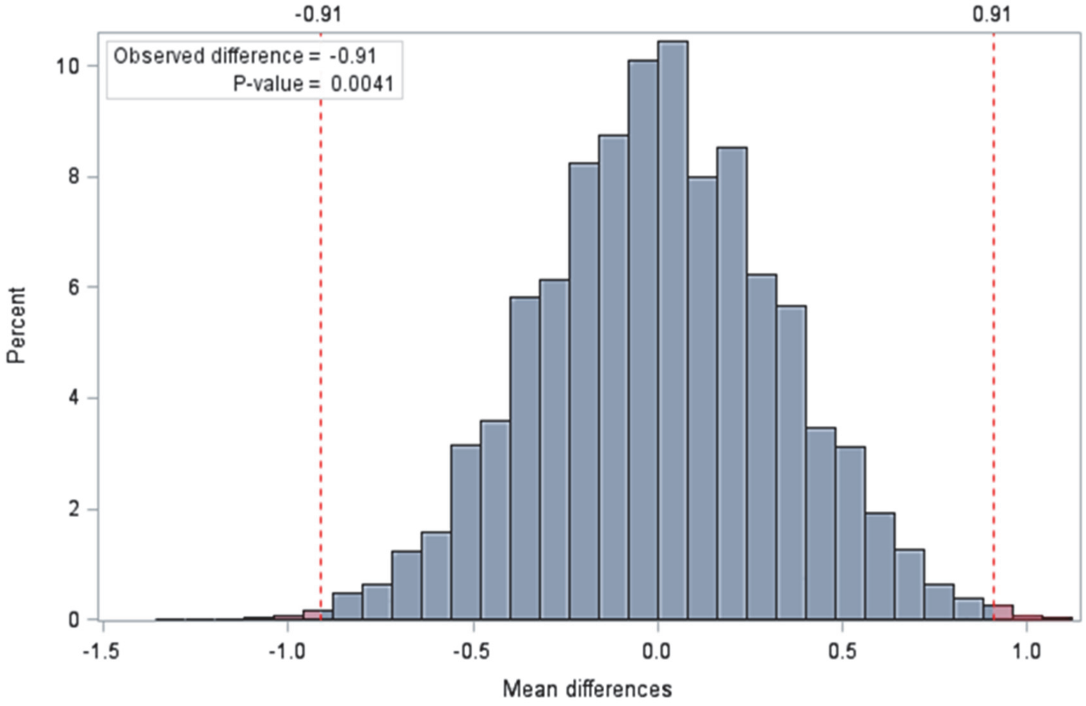


Figure 7. Bootstrap distribution under the null hypothesis for 10,000 resamples shows that the observed difference in the mean number of relevant results between Google Scholar and Summon was significant.