Am. J. Phys., **77**(12), 1112-1117 (2009).

A student's guide to searching the literature using online databases

Casey W. Miller,[1,*] Michelle D. Chabot,[1,†] and Troy C. Messina[2,‡]

[1] *Department of Physics, University of South Florida,*
4202 E. Fowler Avenue, Tampa, Florida 33620 USA

[2] *Department of Physics, Centenary College of Louisiana,*
2911 Centenary Blvd., Shreveport, Louisiana 71104 USA

Abstract

A method is described to empower students to efficiently perform general and literature searches using online resources. The method was tested on undergraduate and graduate students with varying backgrounds with scientific literature. Students involved in this study showed marked improvement in their awareness of how and where to find accurate scientific information.

I. INTRODUCTION

One of the most important tools for researchers is the ability to find and judge the work of other scientists. These talents are developed over time, but can be expedited by a working knowledge of how to efficiently use internet databases. Literature search tutorials provided by libraries[1,2] are too general for students to easily apply to specific databases or disciplines, causing novice researchers to spend large amounts of time performing what are often fruitless searches. To assist students some institutions have implemented courses devoted to science literature searches. A recent study focusing on the health sciences concluded that even university faculty members are unaware of the common tools available for literature searches.[3]

In this article we demonstrate how students can find the most influential papers on a general topic, and then find the most influential papers related to a specific research project. The former will be useful for augmenting students' knowledge base of a topic, which will play an important role in presenting "the big picture," both for original articles and professional presentations. The latter is essential for avoiding duplicating prior work, for determining additional questions to investigate, and for developing a thorough yet concise list of references. For both goals, we describe and demonstrate an algorithm of search, sort, inspect, and repeat.

We assume that students have an initial idea of the appropriate general search terms through discussions with an experienced researcher. We focus on ISI's Web of Science[4] because we have found this database to be the most comprehensive and flexible for physics. Other major databases ought to work just as well,[5] at least for the physical sciences. In medicine it is necessary to use multiple databases to perform comprehensive searches.[6] However, the authors realize that databases such as ISI are usually subscription-only services. Google Scholar[7] and arXiv.org are free databases that allowing access to a diverse and complementary set of articles, but neither has the flexibility of ISI's Web of Science. For example, Google Scholar provides a complement to ISI because it searches patents as well as articles. Two current problems with Google Scholar are that its citations are inaccurate, and it does not allow for advanced sorting as we discussed below. The arXiv.org complements ISI because it offers preprints. Two drawbacks to arXiv.org are that it contains un-refereed material, and does not enable advanced sorting.

II. TOOLS OF THE TRADE

We first need to become acquainted with the basic search parameters for the database we have chosen. For ISI these include (a) search type (general, advanced, cited reference), (b) citation database, (c) time range, and (d) "field tags." Throughout a typical search session, (a)–(c) are fixed, and (d) can change. The most useful combination of these parameters for an active physics researcher is (a) advanced search, (b) SCI-EXPANDED citation database (a filter that eliminates arts and humanities), (c) search all years, and (d) search by topic (`TS`).[8] For Google Scholar, "Advanced Scholar Search" is recommended. To search by topic in the arXiv, use the "Full Record" option.

A flexible database is the key to efficient literature searches, and for this reason the following discussion mainly focuses on ISI Web of Science. The most fundamental concepts for efficient searching are field tags, search string perturbations, and ability to sort the results.

A. Field Tags

Field tags tell the search engine where to look in the database to find entries that compare favorably with the search string. The most useful of these are `TS` (topic), `AU` (author name), `ZP` (zip/postal code), `SO` (journal title or source), and `TI` (article title). These field tags can be combined with boolean operations (`AND`, `OR`, `NOT`) to narrow or expand the search results.

- `TS` is the most frequently used field tag and finds matches to the search string in article titles, abstracts, and keywords.

- `AU` is useful for perusing a specific researcher's body of work.

- `ZP` enables consolidation of search results by postal code, for example, when an author's name is not unique.[9]

  `SO` is used to find articles from a specific journal and is useful when one wants to view articles in the most prestigious journals or a specific field's focus journals.

  `TI` is useful when searching for a single, known article, and sometimes when a `TS` search yields an unmanageable number of hits. We recommend avoiding `TI` searches because many articles are missed because the search string was not used in the article's title, though it may be present in the abstract. As an example `TS=giant magnetoresistance` and `TI=giant magnetoresistance` have 7402 and 1457 hits, respectively.

B. Search Adjustments

Two important tools for adjusting the search string are phrase searching and truncation. Phrase searching finds exact matches to a string enclosed by quotes (for example, "`giant magnetoresistance`"), and is often useful when a specific phrase is common in the field. For example, `TS=giant magnetoresistance` contains a logical `AND` between `giant` and `magnetoresistance`, but does not distinguish results based on word order. Consequently, spurious results such as "giant hysteresis of magnetoresistance" can be avoided by using `TS`="`giant magnetoresistance`." To complement this feature, truncation increases the flexibility of searching, and becomes handy when spelling might exclude some hits. Asterisk truncation allows searching for a string that is followed by any number of additional characters. Truncation is most useful when the string might not be in its plural form (for example, `TS=superlattice*` will find both superlattice and superlattices), or if only the root word is of interest (for example, `TS=magneto*` will find magnetocaloric, magnetoimpedance, magnetoresistance, magnetosphere). Note that hyphenation is not an issue for `TS`, at least with magnetoresistance, which is sometimes spelled magneto-resistance. In cases where one character may or may not be present, as is commonly a problem with British and American spellings, a $ placed in the location of the optional character will yield all hits with and without that character: `TS=quark flavo$r` is equivalent to `TS=(quark AND (flavor OR flavour))`.

Refining and iterating the search algorithm is key. But how do students learn to meaningfully refine their searches? Initially, a more senior researcher will be necessary to guide the student in this respect. It is also important to know a research project's boundary conditions (for example, specific instrumentation may be desirable). Additional perturbations can be discovered by identifying keywords or concepts that appear repeatedly in a subset of the articles. Searching the literature is a skill, and students need to realize that "practice makes perfect."

C. Sorting

How a particular search engine orders results is a key feature of that database. The options available for sorting results vary by database. When a search is performed in ISI Web of Science, the results can be sorted by a number of parameters. Most significantly, results can be sorted by date or by number of citations. Google Scholar automatically orders the results using a complicated algorithm that considers factors such as the date, times cited, author, and journal. There is an option to find "recent articles" in Google Scholar that will increase the importance of the date in this algorithm. There are no other ways to customize the way that Google Scholar sorts results. The results of a search using the arXiv are always sorted by date, without any options for customization.

Sorting by the number of citations is important when trying to locate the articles which are most highly valued in a particular field. By examining the most cited articles in a search result, the key ideas and necessary background information for a topic can be easily obtained. This method does not imply that articles with low citations are necessarily of lesser importance, as is obvious if we consider recent publications that have not had time to be judged by the scientific community. Furthermore, it is possible that a highly refuted article makes it into the top-cited list purely because it purports ideas that are unpopular.

Sorting by date can also be useful to determine the timeliness of the results and to examine those articles that have not yet had a chance to receive citations. ISI will return the latest articles published in peer-reviewed journals. Often, even more recent results are desired, and the arXiv is an invaluable resource for this type of searching, and is the best way to find preprints and relevant articles about ongoing studies. However, because the results returned from the arXiv are not peer-reviewed, searching the arXiv alone is not a sufficient way to research a topic.

D. Advanced Sorting: The h-index

A useful way of sorting a search is by number of citations, which allows inspection using the h-index concept.[10] The h-index is determined as follows. The articles are first ranked by the number of times that they have been cited such that the article ranked 1 has the most citations. Thus, since the citations descend as rank ascends, the rank must exceed the number of citations somewhere on the list; the rank where the crossover occurs is defined as the h-index. As illustrated in Tab. I, a set of articles with $h = 6$ has six articles each with six or more citations. The seventh article in this example has a rank greater than its citations. It may be instructive to see how h can be determined graphically: h is the rank nearest to the intersection of $c(p)$ with the line $c = p$, where $c(p)$ is the number of citations for paper p in the ordered list (Fig. 1).

Applying the h-index to individuals has proven to be very effective. An individual's h-index is found by ranking that individual's articles by times cited, and finding the value at which the rank equals the number of citations. Hirsch showed that successful scientists have large h indices (and more importantly a large value of dh/dt).[10] Although Hirsch presented considerable evidence suggesting that no universal criterion exists for determining whether or not an absolute h-index can be qualified as "good" or not, larger values of h always implies "more influential." A hind-sight study showed a strong correlation between h and committee peer review,[11] and indicates that h measures how one's contributions are viewed by one's peers. Without loss of generality, we may apply the h-index to any collection of articles returned from a search to find a subset of the most influential (most highly cited) articles correlated with the search string.[12] This use of h can be a useful guideline to determine approximately how many articles returned in a search should be examined. For example, we could read the top h articles to sample the most influential work in a specific area. As mentioned, this method does not imply that articles with rank greater than h are of poor quality.

III. EXAMPLE SEARCHES

The basic methodology we employ is search, sort, inspect, and repeat. Searching is performed with the field tags and search string perturbations we have introduced. Sorting is performed to determine the highest ranked articles of a specific search. Inspecting the top-cited articles of a search will lead us to the most influential topics or keywords within the results, which will then allow us to refine the search string, and then repeat these steps.

We will use this methodology to perform two example searches on the topic of multiferroics, one general and one specific. For the latter, we imagine a student having been asked to determine if, what, and how multiferroics have been grown by sputter deposition. A set of articles uncovered by various searches allows us to draw several conclusions about the field in general, and the deposition process of one specific compound. A summary of both the general and specific searches is presented as a flow diagram in Fig. 2, including search strings, number of hits per search, and conclusions drawn from the article sets returned by the database.[14]

The efficiency of this method relies on active researchers' interpretations of previously published work in their field – their citations are an indication of an article's quality. The information we learn about multiferroics gives us a general idea of the important issues in this area without having to read more than titles. More detailed information can be obtained by reading abstracts to identify recurring ideas or topics. Ultimately we will have to read the articles in their entirety to judge their appropriateness.

A. General Search

We begin our general search with `TS=multiferroic`, which yields 504 hits – a lot to read, but less so if 50 hits per page are shown. For completeness, we then adjust to `TS=multifer*` and find 630. If we sort by times cited, we find that the some of the top articles contain the word "multifermentans," which appears in biochemistry journals. These are errant hits, so we adjust the search to `TS=multiferr*` and obtain 586 hits. Of the top-cited articles, the majority are in respectable journals: *Physical Review B*, *Physical Review Letters*, *Science*, *Nature*, *Nature Materials*, *Applied Physics Letters*, and *Journal of Applied Physics*. We notice by perusing the titles that many articles are concerned with bismuth ferrite, $BiFeO_3$. We also notice that most of the top-cited articles have been published within the last five years. We can extract the number of articles published each year, plot the results as shown in Fig. 3 and see that this field is really just starting to take off. (To do this, go to "Analyze Results" and rank the records by publication year, or use the "Create Citation Report" option.) In so doing, we notice that one anomalous result was published in 1996, several years before the majority of multiferroic articles. This article is entitled "Clusters of lymphoma in ferrets," and should clearly be disregarded in the analysis. Near this article, we discover another errant hit published in the *Journal of Coordination Chemistry*, which passed through the filter because it is concerned with "multiferrocenyl groups." At this point, we refine our search systematically and use `multiferro*` (584 hits), `multiferroi*`

(576 hits), `multiferroic*`(573 hits). With `(TS=multiferro*) NOT TS=(multiferroic*)`, we realize that `multiferrocenyl` accounts for the majority of the eleven article difference between these two search strings. Ultimately, we choose `multiferroic*`, which takes care of the plural and singular usages. Alternatively, we could have searched `(TS=multiferro* NOT multiferrocenyl)` to obtain the relevant 573 articles. Having convinced ourselves that this search string contains the most appropriate articles, we now proceed to investigate either the big picture or specific details related to multiferroics.

To get the "big picture," we chose to view articles published in the top journals. *Physical Review Letters* is one of the most prestigious physics journals, so viewing its articles will give us a glimpse of the top research on multiferroics.[13] This search is achieved with the string `TS=multiferroic* and SO=Physical Review Letters`. There are only 27 articles that meet these criteria, making this search an excellent starting point. By sorting by times cited, we notice a variety of compounds in addition to the previously identified $BiFeO_3$: YMn_2O_5, $TbMn_2O_5$, $DyMnO_3$, $EuTiO_3$, $SrTiO_3$, $TbMnO_3$, and $CuFeO_2$. In this manner we have discovered that multiferroics research focuses on compounds containing oxides of Fe, Mn, and Ti. It is likely there are other multiferroic compounds being studied, but this procedure has identified those with the most success at the present time. We can learn more by quickly scanning the titles of these articles for recurring words or concepts. This scan makes it apparent that the following topics are important: frustration, strain, epitaxy, films, and polarization.

For a deeper understanding of current issues we should read the introductory paragraphs of the articles published in *Physical Review Letters*. Because *Physical Review Letters* is intended to have a broad readership, the introductions are more accessible to someone just learning about the field (the body of the article may be another story). Further, because the aim of a good introduction is to describe the background and history of the work presented in the article,[15] any work cited in the introduction should also be read. After reading a dozen or so articles, paying particular attention to their introductions and conclusions, we will begin to uncover the big picture.

B. Specific Search

For specific details we must know more or less what we are looking for, but should begin using broad terms. Starting a search that is too focused will inevitably exclude pertinent articles because different authors choose slightly different words or methods to describe or investigate the same problem. Suppose, for instance, that we are interested in sputter deposition of multiferroics. We should first determine if sputtering has ever been used to grow any multiferroic films, then aim to determine specific materials and deposition parameters (temperature, pressure, or substrate type). `TS=multiferroic* sputtering` yields 7 hits; `TS=multiferroic* sputter*` yields 9 hits; `TS=multiferroic* sput*` yields 9 hits. Now we should be convinced that very few of the 573 multiferroic articles have sputtering as the thin film deposition technique. What method is preferred? A search using `TS=multiferroic* depo*` yields 109 hits. Sorting by times cited, this search has an h of 14. The majority of the top-cited articles use pulsed laser deposition, and only one used sputtering. Because we are concerned with sputtering, we return to the 9 articles on that topic as a starting point. From these titles we discover that $BiFeO_3$ is the most commonly grown material by (reactive) sputtering. We therefore start a new search: `TS=((bifeo* OR bismuth ferrite) AND (reactive OR sput*))` with 38 hits, several of which are in *Applied Physics Letters* and *Journal of Applied Physics*, which are more specialized journals than *Physical Review Letters*. A quick glance at the abstracts reveals that this material has been grown on glass, mica, Si, MgO, and $SrTiO_3$ (using various buffer layers) at temperatures from ambient to 800°C, and pressure is used to tune the stoichiometry and structure. Further, the films are reactively sputtered from off-stoichiometric (Bi-rich) compound targets.

C. Putting it all together

By using the tools and methodology we have presented, we have uncovered general multiferroic references, and specific references related to sputter deposition of bismuth ferrite (see Fig. 2). The general references (and pertinent references contained therein) will assist us in understanding the global issues in multiferroics. The deposition references (and relevant references contained therein) will help determine where we might begin to learn how to grow $BiFeO_3$ by sputtering, what obstacles we might encounter, and what has already been done with this technique. Following a trail of references will lead to additional general and specific information. Usually one to two degrees of separation is sufficient to find closely related articles. For a complete picture that surrounds a specific article, we must inspect both its cited and citing articles.

IV. STUDENT PERSPECTIVES

We have implemented the search procedure we have described in undergraduate and graduate courses. Most graduate students found the methodology to be straightforward, citing previous search experience. They particularly appreciated the field tags and search adjustments, and several remarked that this methodology helped them find articles related to their research projects that they had not previously found. In what follows, reported errors are estimated by standard deviation of the mean.

Twenty undergraduate "Advanced Laboratory" students were surveyed to determine their impressions of the methodology, as well as to quantify its impact. On a scale of 1–5 with 5 being the best, the total perceived usefulness of this methodology was 4.4 ± 0.2, with 60% responding "5." Based on their indications of prior online journal database use, 9 students were identified to have "little or no prior use" (Group A), and 11 were identified as having "some prior use" (Group B). To gauge the impact of this experience, they were asked to indicate what resources they would have used to gather information for science term papers prior to and after learning this methodology (Fig. 4). The options included books, print journals, magazines, Google, Wikipedia, and online journal databases; the scale was 1–5, with 1 indicating "never," and 5 "always." Group A showed a statistically significant change in the use of online journal databases, which jumped from 1.4 ± 0.2 to 4.4 ± 0.2. Of marginal statistical significance for Group A was a decreased mean and increased variance for future use of Google. Of marginal statistical significance for Group B was an increased mean and increased variance for future use of books, and a decreased mean and increased variance of future use of Google. The increased variance implies that students' perception of Google as an appropriate resource decreased. This sentiment is echoed in Fig. 5, which shows the students' perception of where to find reputable science. Group A holds Google and Wikipedia in higher esteem than Group B, and Group B holds books, journals, and magazines in higher esteem than Group A. It is apparent that Group B is more closely reflecting the responses of professional scientists, which suggests that repeated exposure is needed to solidify students' perception of where one is most likely to find reputable scientific results. To this end, we note the significantly lower variance of Group B's rankings of online journal databases. Thus, this endeavor has provided a solid first step for new users, as indicated by the very large increase in likelihood to use online journal databases, and has also helped to reinforce which scientific resources are reputable among more experienced students.

V. SUMMARY

Knowing how to use an online database for literature searches can lead to a significant amount of information in a relatively short period of time if performed in a systematic fashion. The method described here allows one to converge on the desired information in a limited number of searches, whether that information is of a general or specific nature, by using a database that allows sorting by the number of citations. The techniques demonstrated here will help students of any level find influential articles related to a given topic, be it for a term paper or a research project. Students involved in this study showed marked improvement in their awareness of where to find sound scientific information.

Acknowledgments

Work at USF was supported by the National Science Foundation. Work at Centenary College was supported by the Mattie Allen Broyle's Inaugural Year Research Chair and the Gus S. Wortham Endowed Chair of Engineering.

[*] Electronic address: cmiller@cas.usf.edu

[†] Electronic address: mchabot@cas.usf.edu

[‡] Electronic address: tmessina@centenary.edu

[1] University of California Berkeley, "Teaching library internet workshops" , `<www.lib.berkeley.edu/TeachingLib/Guides/Internet/FindInfo.html>`.

[2] D. Hirst, “Conducting a literature search,” <www.mace.manchester.ac.uk/aboutus/informationcentre/informationskills/searchguides/guideforstudents.pdf>.

[3] N. Espinoza, A. Rincon, and B. Chacin, “Use of web-based tools by health sciences professionals at a Venezuelan university for searching scientific literature. A cross-sectional study,” Profesional de La Información **15**, 28–33 (2006).

[4] <www.isiknowledge.com>.

[5] For a supplemental search of the Chemical Abstracts Service (CAS) database via SciFinder Scholar, see <www.centenary.edu/attachments/physics/tmessina/litsearch/literaturesearchsi01.pdf>.

[6] C. G. Smith, A. S. Herzka, J. F. Wenz, and E. P. Henze, “Searching the medical literature,” Clinical Orthopaedics and Related Research **421**, 43–49 (2004).

[7] <www.scholar.google.com>.

[8] Text appearing in typewriter font indicates search strings used in online databases.

[9] Careful attention must be paid to work done at different stages of an author’s career, because academics often change institutions.

[10] J. E. Hirsch, “An index to quantify an individual’s scientific research output,” Proc. Nat. Acad. Sci. USA **102**, 16569–16572 (2005).

[11] L. Bornmann and H.-D. Daniel, “Does the h-index for ranking of scientists really work?,” Scientometrics **65**, 391–392 (2005).

[12] M. G. Banks, “An extension of the Hirsch index: Indexing scientific topics and compounds,” Scientometrics **69**, 161–168 (2006).

[13] If the prestige of a journal is unknown, we can estimate the journal’s overall impact by analyzing its h-index in analogy to the h-index of an individual.

[14] All search results were found in late December, 2007, and will have changed by publication of this article.

[15] Editorial, “The aim of a good introduction,” <prl.aps.org/edannounce/PRLv95i17.html>.

Figure captions

TABLE I: Example determination of the h-index for a collection of articles with $h = 6$. The double line indicates the boundary defining h.

Rank		Times Cited
1	$<$	33
2	$<$	18
3	$<$	15
4	$<$	9
5	$<$	9
6	$\leq$	7
7	$>$	5
8	$>$	5
9	$>$	0

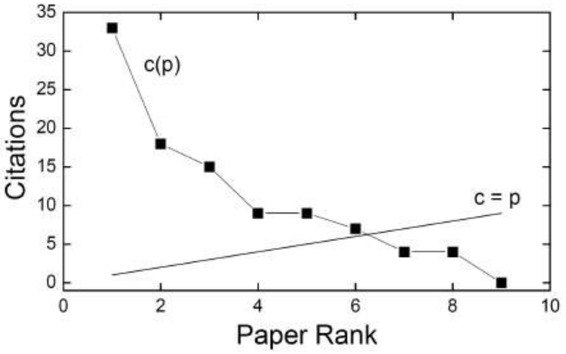

FIG. 1: Example graphical determination of the h-index for a collection of articles with $h = 6$. The rank nearest to the intersection of the line $c = p$ with $c(p)$ indicates h.

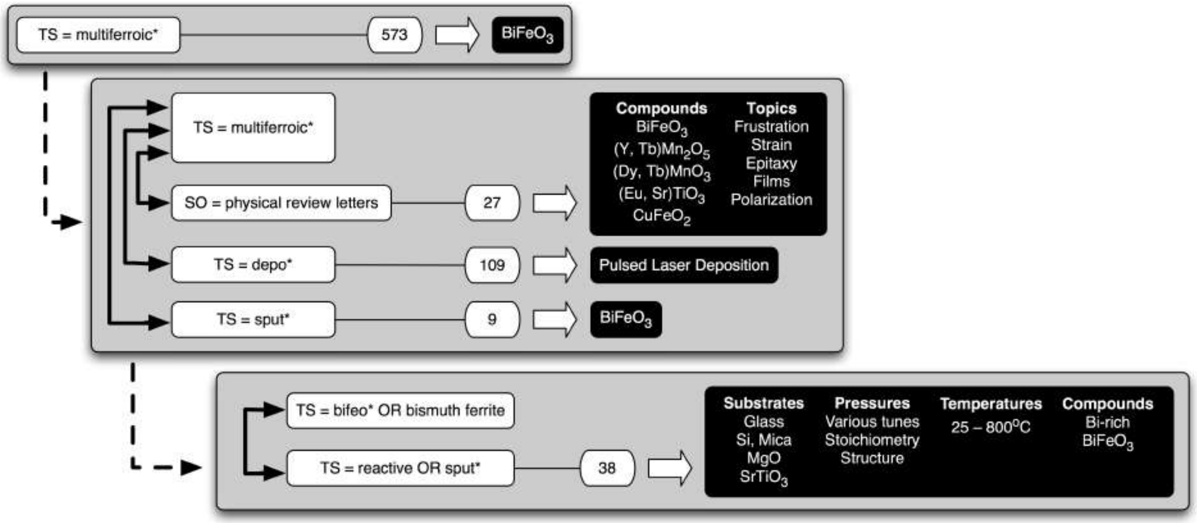

FIG. 2: Mapping the search. Search strings are in white boxes. Double-headed arrows indicate a logical AND. Number of hits from a search are connected with thin lines. Block arrows lead to conclusions (collected in black boxes) drawn by inspecting articles returned from the different branches of the search. Dashed arrows indicate a transition from general to more specific levels of searching.

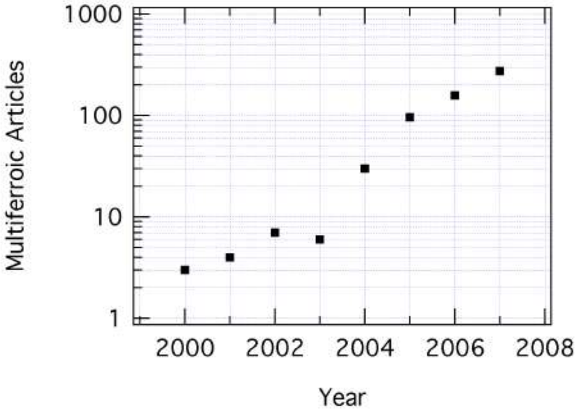

FIG. 3: A plot of the annual number of multiferroic publications versus publication year shows that this topic is just starting, and interest in it is increasing rapidly.

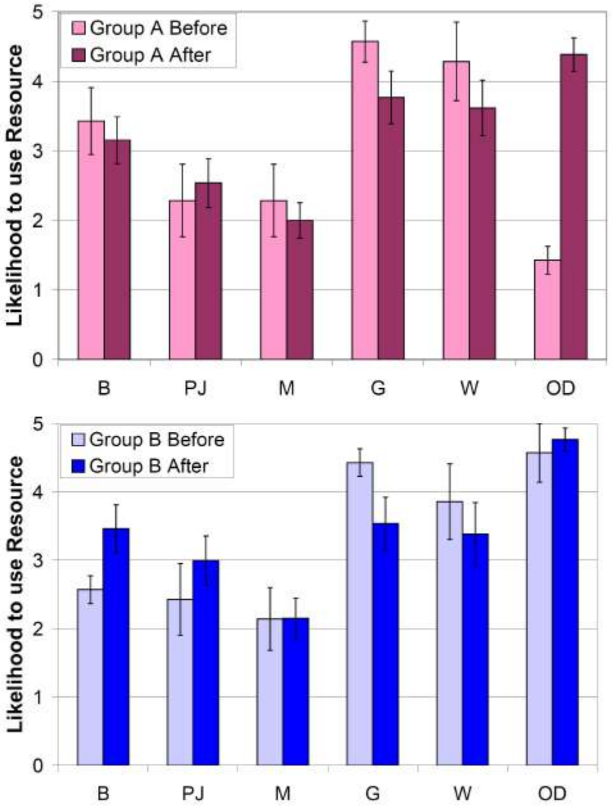

FIG. 4: (Color online) Likelihood of Groups A and B to use available resources before and after our methodology was introduced, based on a 1–5 scale with 1 = “never” and 5 = “always.” The categories were B = books, PJ= print journals, M = magazines, G = Google, W = Wikipedia, and OD = online journal databases.

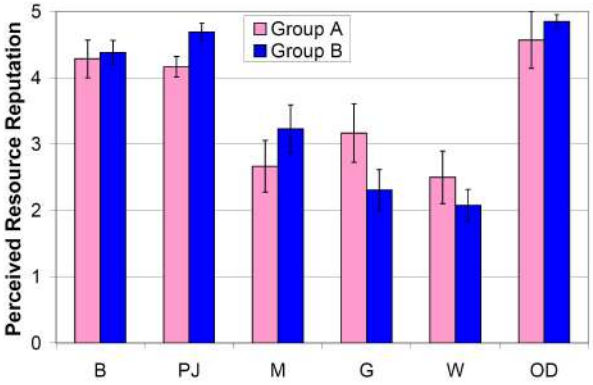

FIG. 5: (Color online) Students' perceptions of the scientific reputation of available resources on a scale of 1–5 with 1 = "least reputable" and 5 = "most reputable." The categories were B = books, PJ= print journals, M = magazines, G = Google, W = Wikipedia, and OD = online journal databases.