Making Hard Choices:
Using Data to Make Collections Decisions

Susan Elizabeth Edwards

University of California, Berkeley

Abstract: Research libraries spend millions of dollars acquiring, storing and accessing collections -- but how well do the collections we build meet the needs of our users? How do we know if we are equitably supporting a wide array of disciplines on campus, and more importantly, how well we support the research mission of our institution? In order to arrive at meaningful answers to these questions, we need to go beyond simple size measures such as dollars spent, volumes added, and number of e-journals licensed; and even beyond usage metrics such as interlibrary loan, total circulation and e-usage statistics. This paper will outline several approaches being used at Berkeley, including a citation analysis of doctoral dissertations, and how this new data is helping guide these difficult decisions.

Keywords: collection development, collections assessment, citation analysis

1. Introduction

The Collections Budget Group at Berkeley (CBG) includes the Associate University Librarian for Collections and fund coordinators from each of the major disciplinary groups (Arts and Humanities, Area Studies, Social Sciences, and Sciences). CBG discusses how to equitably distribute funds within the disciplinary groupings, how to best use discipline-based funds to respond to the growing interdisciplinary and cross-disciplinary nature of research, how to address newer formats such as geospatial data and e-books, and whether we have sacrificed the monograph budget in order to support large e-journal packages with their annual cost increases. Data is readily available to compare how much Berkeley spends relative to other research libraries; how many books we buy and online resources we license -- but that doesn't tell us whether we are

buying the "right" books and journals, or whether we are equally supporting students and faculty in a wide range of disciplines.

There are many different ways to define the quality of a research collection, and these definitions have changed over time. For decades, Berkeley and other research libraries strove to build collections with comprehensive excellence in many languages, obtaining the highest percentage of scholarly output from as many countries as possible. Collections were valued for their size, depth and for the number of unique items included; current need was only one of many factors – in fact, there was an explicit goal of collecting for the patron who would need the material in one hundred years.

Changes in scholarship, academic publishing, information access, the cost of space and the decrease in collections budgets as a percentage of university expenditures, no longer make this goal obtainable -- or perhaps even desirable. But developing consensus on new measures for assessing research level collections has been quite challenging. It's always possible to buy or license more resources -- how do we know what is *adequate* to support a research level collection, and how do we determine an *equitable* distribution? And as individual selectors, how do we know how well we are doing – and how do we define success?

2.   Collection Assessment: the Berkeley Context

Selectors at Berkeley can generate several collection and circulation reports through the Millennium Integrated Library System. These include reports on the age of the collection (by Library of Congress call number) and circulation by patron type (undergraduate, faculty, visiting scholar, etc.) Selectors can also consult a report of all titles borrowed via Interlibrary Loan, a helpful approach for identifying gaps in the collection.

In addition, the Library Systems Office (LSO) has created a number of special reports that are essential for larger scale collection reviews. The LSO reports include the number of circulations an item has received – including how much usage in the past five years --and whether a copy has already been stored in one of the two Regional Library Facilities (RLFs). The RLFs house low use material from all ten campuses. Only non-duplicative material can be stored in an RLF, so these reports are essential in making usage-based decisions of whether to keep a title on campus, transfer it to an RLF, or withdraw the duplicate local copy.

Berkeley is part of a ten campus University of California system, and the California Digital Library (CDL) negotiates large e-journal packages on behalf of the ten campus consortium. To assist selectors in deciding which titles to add or drop from the package, CDL provides a complex array of data for each title

on usage, quality and cost effectiveness. Factors include cost per use, impact factor, and Source Normalized Impact per Paper (SNIP): Wilson and Li (2012).

3. Space Constraints

Like many urban universities, campus space pressures at Berkeley have heightened and the library has not been immune from these pressures. Combined with the increasing reliance on digital resources, decreasing use of the print collection and the high overhead of staffing multiple locations, the university has begun to explore closing or consolidating branch libraries. Table 1 shows the drastic drop in the number of print book circulations (not including renewals or reserves) for each subject library in the social sciences, from 2004-2013. The percentage of decrease for the Education Psychology Library was 68%, the Social Welfare Library had a decrease of 75%.

Table 1
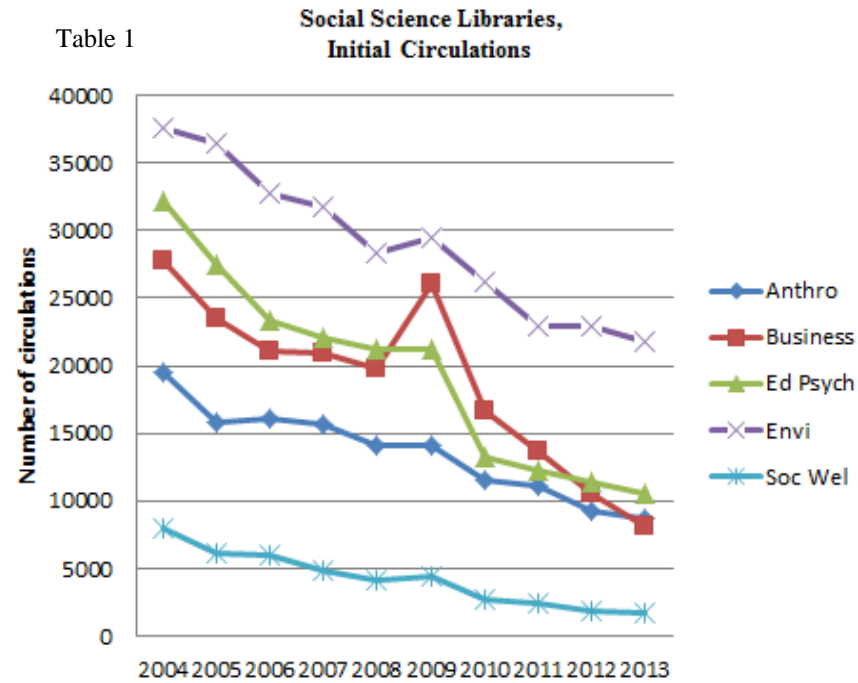
As part of the exploration of library consolidations, the LSO was able to provide new data on circulation by patron's discipline by subject library for faculty and graduate students. These reports use the faculty or graduate students department, which is in the patron address file, for a specialized cross-tabulation.

Unfortunately, they are not available for undergraduate students, but even with that limitation they have been incredibly helpful. These reports give us a more detailed picture of how the graduate students and faculty in each discipline use the print collection of the subject specialty and main libraries. For example, we learned that psychology faculty and graduate students borrow few books, and what they do borrow is primarily from the Education Psychology Library (Table 2); and that education graduate students and faculty check out many more books, and education faculty use the Main Library more than they use the Education Psychology Library (Table 3). A suggestion was made to combine the Education Psychology and Public Health libraries; the circulation by major data showed that there were virtually no overlaps in print usage by the three disciplines and this idea was abandoned. Instead the Social Welfare and the Education Psychology Libraries will be combined, with certain subject areas being transferred to Main and low use items to off-site storage.

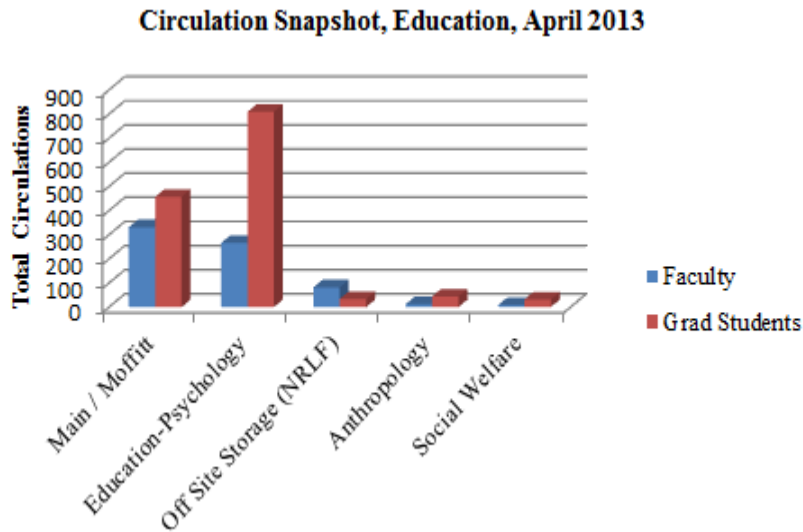Table 2 **Circulation Snapshot, Psychology, April 2013**

Circulation Snapshot, Education, April 2013
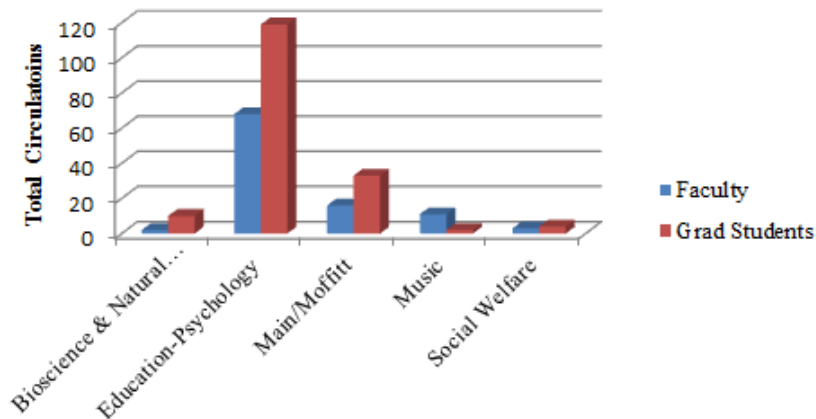
Table 3

4. Comparison With Peer Institution Collections

In 2011, we conducted a comparison of English language print monographs added to Berkeley and Stanford in education and psychology from 2006-2010. I collaborated with my counterpart at Stanford University (our nearest peer research library, and partner in the Research Library Cooperative Program offering expedited borrowing to graduate students and faculty at each of our institutions) to compare our collections. We wanted to determine whether we could lessen duplication and increase reliance on each other in specific subject areas. Both schools' education and psychology programs are doctoral level and English- language based. We used two approaches; a manual comparison by Kathy Kerns at Stanford, and OCLC's WorldCat Collection Analysis (now Collection Evaluation) at Berkeley. Both approaches showed about the same percentage of overlap, and showed a downward trend – from 81% overlap in psychology in 2006 to 53% overlap in 2010; and in education, a decrease from 65% overlap in 2006 to 53% overlap in 2010. Each library analyzed the overlap titles, and the level of duplication was considered appropriate for the research done at each institution. We could not find any benchmarks or best practice recommendations for the ideal level of overlap, so were unable to make any comparisons. However, at a local level we did not see the ability to generate savings through greater shared collection development in these subject areas.

There are many methods of collections assessment: Brown and Stowers (2013) and the data generated in-house, by CDL and by WorldCat Collection Analysis has been an essential tool in collection management – but it has not answered

our central question of how well our collections meet the needs of our users, and whether the disciplines are equally supported.

5. Citation Analysis

As the selector for education, psychology and social welfare, I wanted a metric which would allow me to analyze and compare collection support for each of the three disciplines and ideally to compare this library's support to other libraries of a similar size. The demographics of each department vary (Table 4), but since each has a research-oriented doctoral program a dissertation citation analysis seemed a good choice.

Table 4

Demographics	Faculty FTE	Undergrad Majors	Graduate Students
Education	40	Minor Only	372
Psychology	35	703	109
Social Welfare	23	324	235

Dissertation citation analysis (analyzing citations of dissertations to see the percentage owned or licensed by the institution) is a well-documented bibliometric: Kayongo and Helms (2012). It provides rich data about students' research behavior and about the level of support provided by the collection. Unlike other usage data (circulation, interlibrary loan, or e-usage), it shows not only that a work was used, but that it was *useful*.

In addition to providing a measure of the Library's support of doctoral research, the analysis we conducted (Edwards and Jones (in-press 2014) has also impacted collection development decisions by providing detailed data on the sources students use – for example, books versus journals versus free websites by discipline, or the median age of citations. The methodology used was a systematic sample with a random start, with a confidence interval of 95% (+/- 4%). Bibliographies of doctoral dissertations from three academic departments at the institution were analyzed: Education (2009-2010), Psychology (2009-2010), and Social Welfare (2009-2011). A statistician was consulted to determine statistical significance of the results. The test for ownership was a chi-square test, which is typical for nominal data or dichotomous data.

The research demonstrated that all three disciplines were well and equally supported for doctoral research by the library's journal collections. But for books, we found that we owned a lower percentage than journals overall, and that we owned a statistically significant smaller percentage of books in social welfare compared to either psychology or education (Table 6).

Table 5: Percent of Citations Owned or Licensed		
	Journals	Books
Education	97%	86%
Psychology	99%	87%
Social Welfare	97%	72%

Another interesting finding was the type of source material cited (Table 7).While students in psychology cited primarily journals, students in social welfare cited a fair number of books, making the discrepancy in ownership even more significant.

Table 6: Type of Sources Cited			
	Journals	Books	Web Sources [government documents, etc.]
Education	46%	47%	7%
Psychology	84%	15%	<1%
Social Welfare	59%	33%	8%

Interdisciplinary and cross-disciplinary research is of increasing importance, but both librarians and faculty were very surprised to see the degree to which some disciplines cite journals which "belong to" (i.e., are paid for) other disciplines. Both education and social welfare cited journals from psychology more frequently than the core journals from their own disciplines -- a factor that must be taken into account when determining appropriate funding levels (Table 8).

Table 7
Most Frequently Cited Journals

Education	Psychology	Social Welfare
Journal of Personality and Social Psychology	Neuroimage	Child Development
Child Development	Journal of Personality and Social Psychology	Developmental Psychology
Journal of Educational Psychology	Journal of Neuroscience	Children and Youth Services Review

Developmental Psychology	Nature [(4th)]	American Psychologist [(4th)]
Journal of Research in Science Teaching [(5th)]	Neuropsychologia [(4th)]	Development and Psychopathology [(4th)]
Journal of the Learning Sciences [(5th)]	Nature Neuroscience [(5th)]	Future of Children [(5th)]
Reading Research Quarterly [(5th)]	Neuron [(5th)]	Child Abuse & Neglect [(5th)]
American Educational Research Journal [(6th)]	Science [(5th)])	American Sociological Review [(6th)]
Applied Psychological Measurement [(6th)]	Journal of Cognitive Neuroscience [(8th)]	Child Welfare [(6th)]
Educational Psychologist [(6th)]	Journal of Neurophysiology [(8th)]	Journal of Consulting and Clinical Psychology [(6th)]
	Trends in Cognitive Sciences [(8th)]	Pediatrics [(6th)]

6. Next Steps

The doctoral citation analysis provided such valuable data about user behavior and level of collections support that in 2013 a group of social science librarians at Berkeley applied for, and received, a research grant from the Librarians Association of the University of California to extend to new subject areas. The second phase of the study will include business, economics, history and political science.

In this next phase, we were able to overcome one of the main drawbacks of doctoral citation analysis –that it is extremely time-consuming to gather the citations from each dissertation for analysis. Berkeley dissertations have been submitted only in electronic format since 2009, and unless embargoed, are published open access and via ProQuest's Digital Dissertations. Working with ProQuest and our local Data Center, we were able to import all the citations in the bibliographies of all the published dissertations in our study into a spreadsheet. This saved a tremendous amount of time over the previous study, where each citation was hand-entered. This time we only needed to hand-enter the dissertations that were embargoed, only in print, using footnotes instead of a bibliography, or otherwise not available electronically – a small percentage. It also made it possible to do a language analysis on ALL the citations in the dissertations, not a sample. Citations in dissertations are coded by ProQuest with the language of the citation. Selectors have had an anecdotal sense of the use of

non-English language sources, but with this data we now know that 28% of the History citations were non-English, (and we know which languages are the most commonly cited), 15% of citations in political science and 1.3% of economics. Business had only two non-English language sources cited in total – and both of those had been translated into English!

The study is still in process, but the language findings alone are of value. The Research Library Group (RLG) Conspectus (http://www.loc.gov/acq/devpol/cpc.html) established collecting levels for research libraries, ranging from the lowest of Minimal to the highest of Comprehensive. One of the main differences between the levels of Research and Comprehensive is that Comprehensive collections include source material "in all applicable languages". Knowing which languages comprise the 28% of non-English language sources cited in history dissertations, and that business students are using virtually no non-English language sources, will help us build more targeted collections, in the applicable languages, which better support the needs of doctoral students at Berkeley. We also hope that other institutions (at least those who submit their dissertations to ProQuest), will be able to use our methodology for their own study. One of our early goals – to benchmark Berkeley to similarly sized institutions – was not possible due to a lack of a standardized methodology: Hoffmann & Doucette (2012). We hope this methodology will help make benchmarking possible.

7. Conclusions

There are many ways to define and assess the quality of research collections, but the level of support provided doctoral students is certainly core. Even with the increasing emphasis on access rather than ownership, it is essential that someone – usually a research library – owns the material, and is able to provide it both to the local researchers and to the academy as a whole. In interrelated disciplines, such as those in the social sciences, citation analysis is one of the few ways to compare the level of support provided by the collection. While it is true that one discipline may be less well supported than another because of bad selection decisions -- rather than a lack of funding -- the degree of support provided doctoral students remains an important indicator. Doctoral citation analysis, along with usage data, Interlibrary Loan requests, peer comparisons, and specialized reports such as circulation by graduate students and faculty by library, provides selectors and library managers with essential information about collection strength, collection synergies, funding equity, and the contribution the library makes to the university's research

References

Brown, J., & Stowers, E. (2013). Use of data in collections work: An exploratory survey. *Collection Management*, 143–162.

Edwards,S., & Jones, L (in-press 2014). Assessing the fitness of an academic library for doctoral research. *Evidence Based Library and Information Practice.*

Hoffmann, K., & Doucette, L. (2011). A review of citation analysis methodologies for collection management. *College & Research Libraries*, 73, 321–335.

Kayongo, J., & Helm, C. (2012). Relevance of library collections for graduate student research: A citation analysis study of doctoral dissertations at Notre Dame. *College & Research Libraries*, *73*, *47-67.*

Wilson, Jacqueline, & Li, Chan. (2012). *Calculating scholarly journal value through objective metrics: California Digital Library*. Retrieved from http://www.cdlib.org/cdlinfo/2012/02/13/calculating-scholarly-journal-value-through-objective-metrics/