11-5-2009

Beguiled by Bananas: A Retrospective Study of the Usage and Breadth of Patron vs. Librarian Acquired eBook Collections

Jason S. Price
Claremont University Consortium

John D. McDonald
Claremont University Consortium

Beguiled by Bananas: A retrospective study of the usage & breadth of patron vs. librarian acquired ebook collections

Jason Price, Head of Collections & Acquisitions

John McDonald, Director, Information & Bibliographic Management and Faculty Relations

Claremont Colleges Library, Claremont University Consortium

Charleston Conference 2009

Abstract

Library acquisitions lore contains a cautionary tale of a patron in a demand-driven environment who spent a huge chunk of the library budget on ebooks about bananas. This story and others like it have been used to perpetuate the argument that demand-driven acquisition will result in collections that don't appeal to a broad audience or are otherwise unbalanced. We apply post-acquisition usage data from multiple libraries to test the hypothesis that patron-acquired versus librarian-acquired collections have different usage profiles. In addition, we analyze their subject profiles to evaluate collection breadth and balance. Our results will help libraries to anticipate the effect of adding a demand-driven component to their ebook acquisition strategy.

Bananas tipped the boat: a cautionary tale

The genesis of this study was the sizeable impact of a story that was told during a debate over patron-initiated selection at the Charleston conference in 2008[1]. It goes more or less as follows:

> One of the first ebook vendors to market ebooks to libraries set up an experimental patron-driven purchasing system for a large academic library in Colorado. Soon after its inception, a business professor at the university assigned a paper related to the economics of the banana industry. One or more industrious students found the ebook platform and clicked through to every ebook in the collection matching a search for the keyword 'banana'. Thanks to these students, and an early patron-driven model that lead to ownership after two full text click to-s, the library found itself to be the no-so-proud owner of every banana related ebook in the vendor's collection. Almost certainly, only a select few of these were relevant to the assignment, so the library 'returned' most of these books for a refund.

The ultimate impact of this event on the Colorado library was small—yet the story lives on in library circles as an oft-repeated cautionary tale. Librarians and ebook vendors alike have used it as 'evidence' that patron-driven selection is a bad idea. The influence of this anecdote should not be underestimated—eight years later it has entered the realm of library lore, and is still circulating today. Sales staff of one major ebook vendor used it to justify their company's decision (recently reversed) not to offer a patron-driven pricing model. In addition, the banana legend clearly won converts in last year's patron-initiated purchasing lively lunch debate[1]. There was a major shortcoming in the arguments expressed on both sides, however: neither had data to support their conclusions. Frustrated that an anecdote had won the day, I resolved to rectify this deficiency, posthaste.

As such, our study was designed to address the conclusion that *seems* to flow naturally from the banana book example: i.e. that patron-driven selection inevitably results in purchasing of ebooks that no one (or no one else) is interested in. First we describe the EBL demand-driven system, which was built to ensure that purchases were based on more than browsing or casual interest, and allowed us to distinguish specious post-purchase use from more meaningful varieties. Then, we use post-acquisition ebook usage data from multiple libraries' EBL collections to address the hypothesis that patron-selected collections are inferior to librarian-selected collections. Finally, we discuss key questions that arose during our presentation.

Terminology & Background

The library industry has used various terms to describe use-based ebook acquisition: from patron-initiated to patron-driven to demand-driven to user-driven. Although we billed our talk as comparing patron-acquired versus librarian-acquired collections, we quickly realized that the nature of these

[1] Polanka et. al. Tossing Traditional Collection Development Practices for Patron Initiated Purchasing: A Debate. Charleston Conference 2008.

collections was not quite that clear cut. Some 'patron-acquired' purchases were certainly made by EBL users that were librarians, and some 'librarian-acquired' purchases were undoubtedly made on behalf of faculty patrons who selected them for teaching or research. The most accurate terms we could divine to describe these collections were user-selected and pre-selected. These terms more precisely reflect the fact that the **user-selected** ('patron-acquired') books were purchased *AFTER* they were used one or more times by a local library user, whereas the **pre-selected** ('librarian-acquired') books were purchased *BEFORE* they were used. Henceforth we use these two terms to refer to the 'patron-acquired' and 'librarian acquired' books or collections.

EBL-hosted ebooks were used for this study because they have the most sophisticated demand-driven system currently available, both in terms of purchase triggers and usage reporting. Key features of this system include:

1. **There is a browse period for every use of every book**
    a. This allows browsing/evaluative use to be separated from meaningful use – when a user clicks into a page of an ebook, s/he may decide it is not useful—such usage is reported separately, does NOT count toward purchase, and was **ignored** for this study (whether it occurred before or after purchase)
    b. The browse period results in a recorded use when content in the book is copied, printed, downloaded, or when a user clicks through to keep the page open for more than 5-10 minutes
    c. As a result, every use we included in the analysis indicated meaningful, post-purchase use (Demand-driven models that lack this feature can be likened to a physical bookstore suggesting that if you touch a book, you've bought it!)
2. **Usage-Type categories for every transaction** made it easy to eliminate pre-purchase use—ALL of the usage analyzed for this study occurred *after* the book had been purchased.
3. **Unique but persistent user IDs for every transaction** available from a proxy-based user authentication system made it possible to separate repeat use by the same person (even over a period of months or years) from use by a variety of people, enabling assessment of the breadth of the audience for each book within each institution, as well as its depth in terms of total number of meaningful uses
4. **Detailed electronic invoice data** available for every purchase transaction distinguish between user-selected and pre-selected purchases and provide an exact date of purchase, so that usage could be analyzed in terms of uses per time to account for the fact that some books were owned much longer than others an thus had much greater opportunity for post-purchase usage.

Specifically, for those familiar with EBL terminology, we treated instances of 'read online' and 'download' usage as equivalent, and ignored all 'browse' usage. This ability to distinguish and ignore casual usage distinguishes the EBL system and this study from all other e-resource usage evaluation that we know of, and may be a major explanatory factor of its success as a demand-driven system.

We used data from the EBL system to address each of the following questions in turn:

1) Are user-selected ebooks used less often than pre-selected ebooks?
2) Do user-selected ebooks have a narrower audience than pre-selected ebooks?
3) Are user-selected ebooks less likely to be used than pre-selected ebooks?
4) Are user-selected collections less comprehensive (or more skewed) than pre-selected collections?

Scope and Design

The Claremont Colleges Library has not purchased any books on the EBL platform—user-selected or otherwise: the data used in this study were from other libraries. EBL was willing to share data from eleven libraries on the condition that their specific identities remain unknown to us[2]. These datasets included 28,327 books purchased over a four year period, 2006-2009 (Table 1). In the aggregate, portions of these owned books were 'used' (downloaded, printed, copied, or browsed for more than 5 minutes) nearly a quarter of a million times (213,887) during this period.

Table 1: EBL data available, 11 Libraries

Library	Purchasing Type	User-selected	Pre-selected	Usage - download	Usage - read online
A	**MIX**	**1131**	**552**	**6773**	**9888**
B	**MIX**	**5246**	**2612**	**42880**	**38329**
C	USER	2198	102	0	11801
D	USER	3010	48	697	15126
E	**MIX**	**4159**	**909**	**17396**	**25604**
F	PRE	0	1451	4905	3082
G	PRE	31	2154	7001	4459
H	USER	801	0	556	415
I	**MIX**	**305**	**336**	**3334**	**2568**
J	USER	2799	53	5	13349
K	**MIX**	**147**	**276**	**2436**	**2283**
TOTAL		**19,831**	**8,496**	**85,983**	**126,904**

Purchasing type varied greatly among the eleven libraries. We divided them into three categories (user-selected, pre-selected or mixed) based on the total number of books purchased each way by each library (Table 1). We selected only the mixed type libraries for this initial study to allow us take differences in user communities into account. This approach allowed us to compare user- vs. pre- selected collections within each library as well as overall. (Initial exploratory analysis showed that overall results including data from all eleven libraries were consistent those from our subsample; data not shown). Thus the results we present are from five libraries: A, B, E, I and K--including 15,673 ebooks that were used 151,491 times (Table 1, in **bold**). Within this sample, the libraries bought and average of twice as many user-selected books, but this does not affect the results because comparisons were made on an average-per-book basis and all 10 selection sets had a large enough sample ($n \geq 147$) to avoid sampling bias.

EBL's model allows for a great deal of customization of demand-driven (user-selected) purchasing. Libraries can choose the number of loans before purchase, the range of books available for purchase (by date, publisher, subject, price range, etc.), and can even choose to approve some or all user-selected

[2] We are greatly indebted to Alison Morin, Kari Paulson and Sally TerBeck of EBL, who provided the datasets that met our criteria and engaged in a number of detailed discussions to help us understand and take into account all of the implications and nuances that affect the interpretation of these data

purchases before they are made. All five libraries' EBL platforms were customized in one or more of these ways to varying extents at different points throughout the study period. In spite of this variation, our results show that differences proved to be extraordinarily consistent across these five libraries. They are consistent despite the underlying variation in the type, degree, and duration of user-selected purchase model customization. We feel this adds to the weight to the evidence we present, rather than detracting from it.

Since these sets contain books that were bought continuously over the 4-year period, books within each set had greatly varying opportunity for post-purchase usage (Table 2). Exploratory analysis showed that purchase date had a statistically significant effect on post-purchase usage and number of unique users (data not shown). For simplicity and ease of interpretation, we designed our usage metrics to take length of ownership of each book into account, rather than including it as an additional causal (independent) variable. Total usage was measured as uses per year owned and breadth was measured in unique users per year owned. Because we knew the exact date that each book was bought, we could calculate usage and users per day X 365 with no error in the denominator, avoiding the hazard of introducing an extra source of error due to a ratio-based dependent variable.

In order to minimize the chance of over-estimating the effect of usage of books owned for a short time, we explored exclusion of books owned for less than six months and less than one year. For instance, if a book owned for just 1 day was used 5 times that day, multiplying those uses by 365 days would drastically inflate its estimated usage per year). Neither the direction of the differences nor the significance of the model changed with either of these subsamples, but the differences observed were less extreme when books owned six months or less (n=2350) were removed (data not shown). No additional material change was observed when books owned 6 to 12 months were also excluded. As such, in order to err on the conservative side, we excluded all books owned less than 183 days (6 months) from the analysis.

Table 2: Average Users, Usage, and Availability period

Library	Ebooks	Average Unique Users	Average Usage	Average Days Available
A	1683	5.63	9.90	475.28
B	7859	5.69	10.09	512.33
E	5068	5.00	8.49	440.30
I	641	4.56	9.21	766.38
K	423	5.28	11.13	784.81

Analysis

As outlined earlier, we were interested in understanding if user-selected collections would have the same depth, breadth, initial use and subject distribution as pre-selected collections. To test the first two questions, we developed a model that included three predictor variables: the purchase type, the library, and an interaction effect between purchase type and library. Our analysis of variance (ANOVA) tests addressed the significance of these predictors on two response variables: usage per year and unique users per year of each ebook. These variables address the extent and breadth of usage of each collection. The latter two questions were addressed more informally, based on observed differences in the patterns of non-use and LC class distribution.

Table 3: Usage, Mixed Purchase Type Libraries

Purchase type	Library	Mean	Std. Dev	Books
User-selected ebooks	A	10.12	13.95	778
	B	8.39	19.31	4723
	E	8.14	15.25	3111
	I	14.36	42.49	277
	K	7.09	13.16	147
	Total	8.61	18.74	9036
Pre-selected ebooks	A	3.67	5.89	498
	B	4.60	11.55	2475
	E	4.61	18.55	766
	I	3.21	8.69	331
	K	3.08	5.47	217
	Total	4.31	12.25	4287
Total	A	7.60	11.92	1276
	B	7.09	17.14	7198
	E	7.44	16.02	3877
	I	8.29	29.88	608
	K	4.70	9.56	364
	Total	7.23	17.04	13323

Results

Total post-purchase usage

Descriptive statistics for the overall dataset (Table 3) indicate that, on average, user-selected ebooks were used twice as often as pre-selected ebooks (8.6 vs. 4.3 times per year). Individual libraries showed 1.75 to 4.5 times higher use of user-selected ebooks. Additionally, user selected collections had larger standard deviations, indicating that the most heavily used books were used very frequently, compared to pre-selected books that had a smaller range of average usage.

Results of the ANOVA for uses per book per year (Table 4) indicated that while purchase type was a significant predictor, library was not. In addition, the interaction effect of purchase type and library had the highest observed power, suggesting that the extent of the difference differed between libraries (see Fig. 1).

Further analysis of the dataset indicated that average values of usage per year were higher for user-selected ebooks than for pre-selected ebooks across all five libraries (Fig. 2). Four of the five libraries had significantly higher means at the 95% confidence interval. Based on these results, we are confident that user-selected ebooks are not less likely to be used than pre-selected ebooks, but in fact, are used at a significantly higher rate.

Table 4: ANOVA Results, Mixed Purchase Model Libraries

Source		Type III Sum of Squares	Mean Square	F	Sig.	Partial Eta Squared	Observed Power[a]
Intercept	Hypothesis	191873	191872	295.056	.000	.979	1.000
	Error	4207	650				
Purchase Type	Hypothesis	35509	35508	26.796	.004	.843	.982
	Error	6615	1325				
Library	Hypothesis	3984	996	.431	.782	.301	.088
	Error	9243	2310				
Purchase * Library	Hypothesis	9243	2310	8.092	.000	.002	.998
	Error	3801351	285				

Figure 1: Purchase type by library interaction

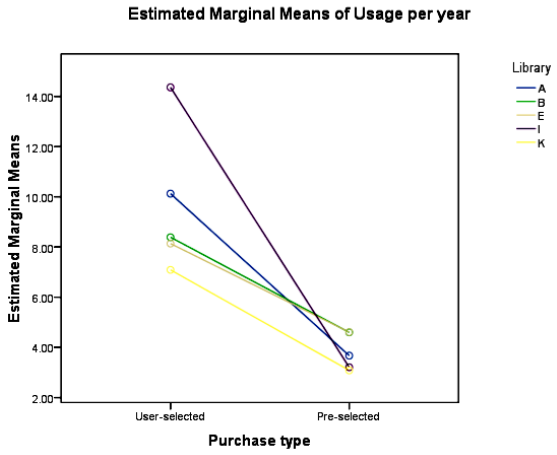
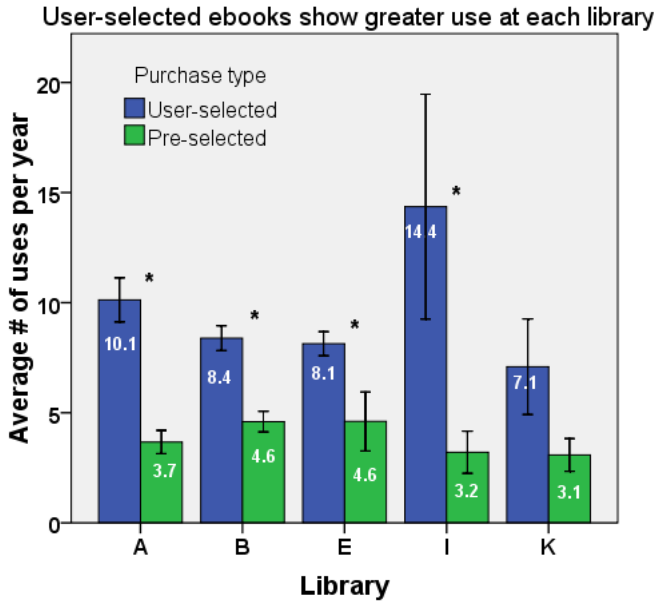
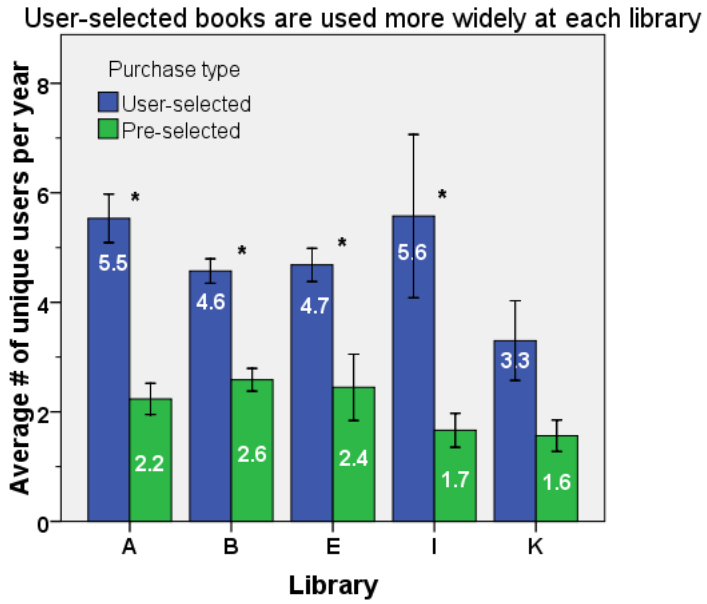


Figure 2: Average usage by library and purchase type



Cell n > 150; Error Bars: +/- 2 SE; * = non-overlapping 95% C.I.

Figure 3: Unique users per ebook per year

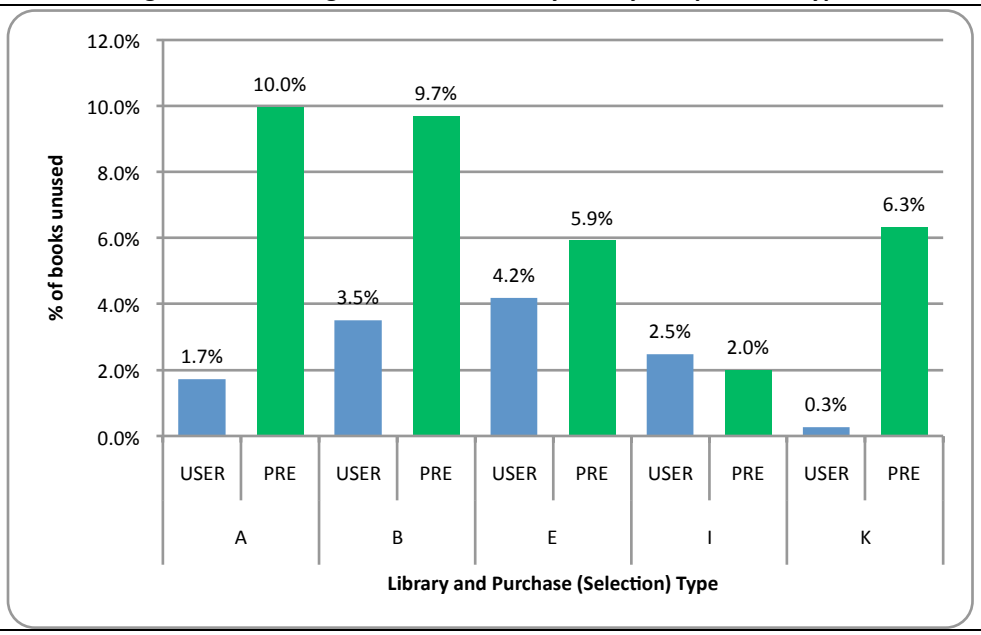


Cell n > 150; Error Bars: +/- 2 SE; * = Non-overlapping 95% C.I.

Breadth of post-purchase usage

While user-selected ebooks are used more frequently than pre-selected ebooks, one could reasonably assert that this is due to the fact that if a user selects an ebook, they are certainly more likely to use it a second time and continue to use that item. Stated in another way, user-selected collections could end up being built in a narrower sense, with each book appealing only to the user who selected it. This observation led to question 2: Do user-selected ebooks have a narrower audience than pre-selected ebooks? On the contrary, the average number of unique users per year was 1.75 to 3.3 times higher for the user-selected collections (Fig. 3). As was the case with overall usage, the average for the user-selected collections was significantly greater than for the pre-selected collection at four of the five libraries at the 95% confidence interval. The ANOVA results (not shown) were similar to the results testing for an effect on overall usage (presented above).

Figure 4: Percentage of titles unused by library and purchase type



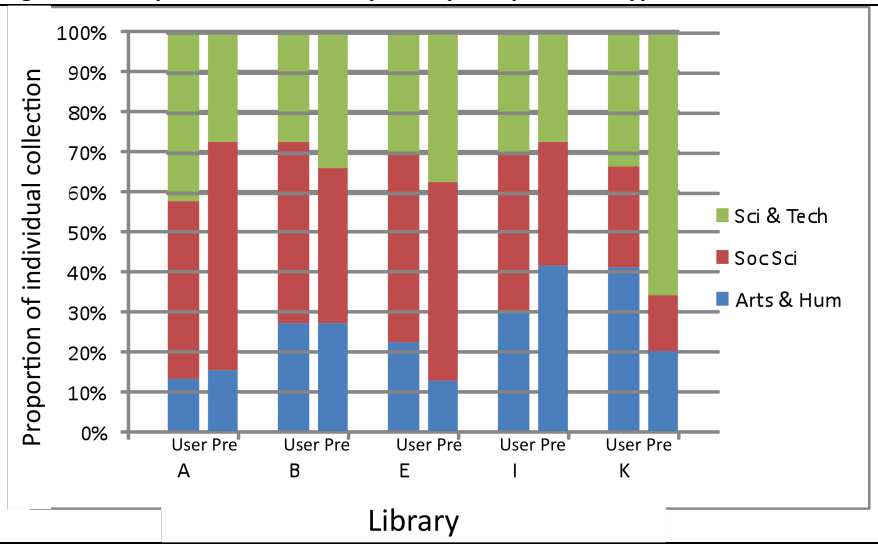
Number of unused titles

While we found that the average number of users per book was greater for user-selected ebooks than pre-selected ebooks, the possibility remained that user-selection could result in a larger number of titles that went unused after purchase. This led to question 3: Are user-selected books less likely to be used than pre-selected books? We looked at descriptive statistics to answer this question. The percentage of unused books was quite low overall: 90% of the books in every collection had been used at least once since they were purchased (Fig. 4). Four of the five libraries appeared to have fewer unused titles in their user-selected collections. Because this is a library-level statistic, however, our sample size (5) was not large enough to test for significance.

Subject coverage

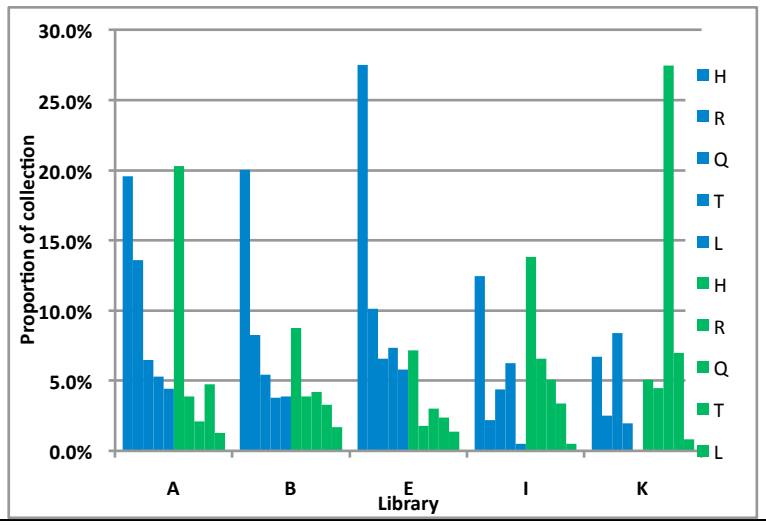
While user-selected collections are used more frequently than pre-selected collections and used by more users (apparently with fewer unused titles), an argument can be made that there is an additional value to the overall library's collection from pre-selected collections in the form of a more balanced collection across all subject areas and research interests for the community as a whole. Our final question addressed subject balance: "Are user-selected collections less comprehensive (or more skewed) than pre-selected collections?"

Figure 5: Discipline distribution by library and purchase type



Discipline level subject distribution appeared to be remarkably consistent across purchase types (Fig. 5). The only distribution that is markedly different is for Library "K", where users selected twice as many Arts & Humanities and Social Sciences books, while the pre-selected ebooks fell more heavily into Science & Technology. This may have been by design, or because this library expected its EBL users to be more science-centric than they proved to be.

Figure 6: Collection distribution by LC class, library and purchase type



Addressing the same question at a more specific level, we categorized the collections by LC class (Fig. 6). The relatively similar distributions of the most common classes of the Library A, B and E collections suggest that the subject distribution of user-selected and pre-selected collections is similar in these libraries. Libraries "I" and "K" had somewhat dissimilar user- vs. pre-selected collection patterns. Library "I" had an almost inverse collecting relationship for some subjects, and library K pre-selected a much higher proportion of 'Q' books as noted above.

This analysis provided good evidence that user-selected collections are no more narrow, skewed, or individually focused than those chosen by a pre-selection. And in fact, for most institutions in the study, the collecting pattern of users mirrored those of pre-selection.

Discussion

These results clearly and repeatedly demonstrate that EBL's demand-driven acquisition model builds collections of ebooks that are used more often and have a wider audience than their pre-selected counterparts. As such, ownership of obscure, unwanted books is *NOT* an inevitable outcome of use-based ebook selection, as the patron-driven banana book legend seems to imply. Furthermore, we present preliminary analyses that suggest that gaining this use-based advantage does not require libraries to sacrifice subject breadth.

The usage and breadth advantage conferred by user-selected acquisition appears to exist in every library we tested. Library K was the only one that did not show a statistically significant effect of purchase type, probably because it had the smallest sample sizes. The broad agreement in this advantage across libraries suggests that it does not depend on the degree of customization of the demand-driven model. Libraries user-selected collections had greater use by a broader range of users regardless of whether they mediated the purchases, or restricted the catalog of books available for purchase, or had a higher threshold of initial use before purchase, or bought their pre-selected books up front or during the entire duration of the study. This consistency should assure libraries that are concerned about the initial set up of their plan that it will have a minimal affect on the demand driven usage advantage.

For some, higher post-purchase usage is sufficient reason to implement a demand-driven model. For others, questions of collection quality remain. Does greater popularity necessarily mean that the collection is better? Will it result in purchases the library would not otherwise have made? Future work could use independent book profile definitions to compare collection quality and address these questions. YBP, for instance, profiles books as Basic Essential/Recommended, Research Essential/Recommended, Specialized, Supplementary, or Unlisted. Comparison of the proportion of books in each category for both purchase types would be enlightening. Publisher and cost breakdowns would also be of interest.

Another important question is whether the higher usage and larger audience we show for EBL-based demand-driven collections would also occur with other demand-driven platforms. More specifically, is the ability to ignore casual click-to use critical to the success of user-selected collections? This question could be addressed within the EBL data by comparing the real life demand-driven collections with those that would have resulted had the browse usage been included. Alternatively, a comparative analysis between patron-driven collections on multiple platforms (i.e. EBL vs. MyILibrary or NetLibrary) could also provide insight. Either of these approaches would have significant methodological challenges, but it seems crucial to take the next step to determine whether a more sophisticated platform is essential for success in demand-driven ebook acquisition.