DIGITAL EDITIONS: MODELING REACH AND FREQUENCY

Julian Baim, Martin Frankel, James Collins, Joseph Agresti, **Seth Cohen,** GfK MRI,
Britta Cleveland, Meredith Corp.
Marlene Greenfield, Hearst Corp.
Caryn Klein, Time Inc.
Scott McDonald, Condé Nast Publications

Introduction

The advent of print digital editions has created new opportunities and challenges for magazine publishers and magazine research companies, alike. For publishing companies, digital editions increase the potential reach of their brands with rich, immersive and interactive content. They afford magazines the opportunity to compete with other media on a more even playing field and to deliver circulation efficiently. At the same time, agencies and advertisers are asking that research companies provide reach and frequency data for the combined print and digital audiences, especially if magazines include their digital edition circulation in rate base guarantees. GfK MRI has already received this particular request from some of the most prominent agencies and advertisers in the United States.

The challenge to create campaign schedules invites a particularly important question: how can we model audience turnover and reach accumulation for digital editions? In particular, what supportive data can be used to model an individual digital magazine's reach and frequency over multiple issues? Can we use the generally accepted beta-binomial model to estimate R&F? And, can we use passively tagged data for unique digital edition openings for these calculations? This paper addresses these questions and provides a justification for integrating print and digital edition R&F calculations.

Background

Over the years, many statisticians have developed models for campaign reach and frequency across multiple titles. All of these eminent researchers have built somewhat unique algorithms for their programs; however, most still base the fundamental individual publication reach curves algorithmic implementation over issues on the seminal beta binomial model work by G.P. Hyett over 55 years ago (Hyett, 1958). Hyett's work was based upon multiple responses from a diary panel to three successive measurements of newspaper reading. His findings provided the empirical verification of the beta-binomial distribution to fit reach and frequency for single publications.

Although Hyett's findings have served as the foundation of R&F campaign estimations, the use of the beta-binomial model from survey data has not gone unquestioned over the years. While accepting the basic fitness of the modeler's formulas, Erwin Ephron noted that "these formulas are generalizations from a limited number of observations so the 'fit of the curve 'is never perfect. And then there's the sampling error of the survey data to which the formula is applied." (Ephron, 1992) A more recent article reaffirms the modeler's quandary of using a limited number of observations:

> The question remains. How can we measure (or estimate) the contact frequency in a target group exposed to a media campaign with multiple insertions of an ad? As it is impossible to collect data from the same respondents for a large number of issues of a publication or different publications, we need models to estimate exposure distributions that are outside the range available in the survey data. (Smit and Neuens, 2011)

The measurement of digital edition readership holds the promise of removing the inherent limitations of survey data. Publishers "tag" their digital editions to allow passive data capture of unique issue openings along with a number of other metrics associated with edition readership. Even though the use of passively captured data is not without measurement issues (see the Mattlin and Gagen paper at this forum), there is strong consensus of the superiority of census-derived passive data over sample survey-based data. Passive data essentially remove non-sampling error from the equation; they also replace the limitations of a sample with a full census. Finally, passive data allow for an unlimited number of observations, uncontaminated by prior survey responses, across time, thereby providing modelers with more robust data for their work.

Given the immediate industry need to justify using the beta-binomial model (or even another model) to develop R&F curves for digital editions and the availability of digital edition passive data, GfK MRI partnered with four publishing companies to provide their Adobe data for a set of their respective magazines. Conde Nast, Hearst, Meredith and Time Inc. (they are listed in alphabetical order for convenience) all agreed to provide digital edition tagged data to help answer the question about the appropriate model for digital issue R&F. Their willingness to participate in this research project enabled us to evaluate the beta-binomial model with passive, rather than survey data. As an additional heuristic research exercise, these companies also provided samples of subscribers and print authenticators[1] (where applicable) to evaluate the beta-binomial model from on-line surveys conducted by GfK MRI.

[1] Authenticators are print subscribers who request and are granted access to the digital edition as well.

Methodologies

All four publication companies embed an electronic code in their digital editions that capture, among a number of metrics, unique openings[2] to individual issues. A number of companies capture these types of data; this study uses Adobe data in our analysis. Each of the companies extracted data over four consecutive digital editions (4 months for monthly publications and 4 weeks for weeklies) from the Adobe database. [3] The data provided counts of unique openings (tantamount to reading or looking into an issue) for the following individual and combinations of issues:

- Total openings for each issue
- Total unduplicated net openings for each of the 6 pairs of issues from the total of 4 consecutive issues
- Total unduplicated net openings for each of the 4 triples of issues from the total of 4 consecutive issues
- Total unduplicated net openings any of the 4 consecutive issues
- From among those who opened 1 or more of the 4 issues:
    - The total of those who only opened issue 1 and not issues 2-4
    - The total of those who only opened issue 2 and not issues 1, 3 or 4
    - The total of those who only opened issue 3and not issues 1, 2 or 4
    - The total of those who only opened issue 4 and not issues 1-3
    - The total of those who opened issues 1 and 2, but not issues 3 or 4
    - The total of those who opened issues 1 and 3, but not issues 2 or 4
    - The total of those who opened issues 1 and 4, but not issues 2 or 3
    - The total of those who opened issues 2 and 3 but not issues 1 or 4
    - The total of those who opened issues 2 and 4 but not issues 1 or 3
    - The total of those who opened issues 3 and 4 but not issues 1 or 2
    - The total of those who opened issues 1, 2 and 3 but not issue 4
    - The total of those who opened issues 1, 2 and 4 but not issue 3
    - The total of those who opened issues 1, 3 and 4 but not issue 2
    - The total of those who opened issues 2, 3 and 4 but not issue 1
    - The total of those who opened all 4 issues (1, 2, 3 and 4)

From these data, GfK MRI tabulated an average issue audience, a 2-issue cumulative audience, a 3–issue cumulative audience and a 4-issue cumulative audience. We also calculated the frequency distribution (i.e.., read 1 of 4, 2 of 4, 3 of 4, or 4 of 4) from among the 4-issue cumulative audience. These figures provided census-level passive data against which we fit the beta-binomial model. Specifically, we input the passively captured average-issue audience estimate and the 2-issue cumulative audience and used the beta-binomial to project the 3-issue cumulative audience, 4-issue cumulative audience and frequency distribution of the 4-issue cumulative audience. We then compared these estimates to the actual passive data, thereby providing an empirical validation of the beta-binomial model.[4]

GfK MRI and the publishing companies shared responsibilities for the on-line survey component of the research project. We conducted the study for exactly the same titles for which we sought to extract the Adobe data. The study consisted of 27 separate surveys of digital subscribers, authenticators and print only subscribers for individual titles. (Results of the print only subscribers are not used in this paper.) All 4 companies generated their respective samples independent of GfK MRI. Two companies e-mailed invitations to their subscribers and/or authenticators while GfK MRI handled the invitations for the remaining two companies. GfK MRI hosted the surveys and completed all tabulations. All surveys were conducted between June and August 2013.

The survey questionnaire contained questions about reading/looking into the 6 most recent issues of a monthly magazine or the 8 most recent issues of a weekly magazine. In every case, respondents were shown cover reproductions of the print and digital issues, respectively, to aid their recall. Other questions asked about frequency of reading, sharing of the digital edition and basic demographic information. Using the 4 oldest issues for each magazine, we tabulated the same individual and combined levels used in our analysis of the passive data. We performed a similar validation test for the beta binomial model against the survey data.

[2] At present, the standard for classifying "opening" of an issue can vary somewhat by publishing company
[3] Due to the complexity of the Adobe database, not all companies were able to provide the requested data in time for this paper
[4] In total, we compared modeled and empirical reach estimates for 21 publications and frequency distribution estimates for 17 titles

Data Analysis and Findings

Evaluating the appropriateness of the beta-binomial model entailed comparing passive data calculating reach and frequency to its model-based counterpart. Specifically, we compared the 3 (C3) and 4 (C4) issue reach figures and the number of issues read among the cumulative 4-issue reach of reading for the empirical passive data and beta-binomial modeled estimates, respectively.[5]

Figures1-11 (see appendix A) show the 3 and 4 issue reach comparisons for about half of the total of 21 comparisons.[6] Since 1 and 2 issue reach figures are required to model additional issue reach, these numbers are exactly the same for the passive and modeled lines. The closeness of fit is demonstrated by the virtual overlap of reach and frequency estimates for 3 and 4 issues. From among the total of 21 comparisons[7], 15 of the C3 modeled estimates were within + or – 1% of the reported C3 audience from the passively tagged data and all are within +/-5%. The analogous comparison for C4 shows 9 modeled estimates to be within +/- 1% of the comparable passive figure and all are within +/-5%. The overall index of modeled to empirical data for the 21 comparison was 100 for C3 and 99 for C4!

The analysis of modeled frequency of reading distributions from among any reading of 4 issues against the passive figures was conducted for available data from 17 magazines. Figures 12-20 (see appendix B) show 9 of these evaluations. Every example shows exceptional consistency between modeled and empirical data. We also noted that any differences between the modeled and passive data in a particular frequency (e.g., 1 of 4) were generally offset by a counterbalancing difference in the adjacent frequency (e.g., 2 of 4).

A second part of the research plan was to validate the beta-binomial model against survey, rather than passive, data. We limited our examination to comparisons between C3 and C4 reach estimates from survey and modeled data for the same magazines. Figures 21-31 once again show these comparisons for the same subset of 11 magazines used in the passive data analysis. Once again, the modeled curves closely parallel the survey curves. All modeled estimates were within +- 3% of the survey C3 estimates with an overall index of 101 of modeled to survey data for 27 magazines. All but one of the 27 C4 modeled estimates were within +/- 5%; the overall index of modeled to survey data was 102.

While the beta-binomial model worked exceptionally well for both passive and survey data, we should note that respective turnover levels from these two data sources were different for each of the magazines in the study. In every case, survey data generated lower issue-to-issue turnover than the comparable passive data figures. There are a number of factors that could explain these differences; among them are the very low response rates from the on-line surveys (below 10%) and potential respondent confusion between digital and paper edition reading. Future research should match respondents' survey responses with their passive data. This type of analysis, currently unavailable from Adobe or the publisher databases, would provide many answers to survey-based magazine audience measurement debates that have preoccupied researchers since the first Worldwide Readership Research Symposium.

Conclusion and Next Steps

The introduction of digital editions has created a rapidly changing, exciting, and challenging environment for print brands. Publishers are not only in the midst of formulating or refining their business model for the digital platform; they are also challenged to provide the necessary metrics for agencies and advertisers to evaluate the effectiveness and efficiency of these editions. The urgency to respond to the print industry is also felt by research companies. In particular, researchers are asked to provide R&F models for digital editions as quickly as possible. The need to respond immediately, however, is offset by the availability of almost real-time passive readership data, thereby enabling researchers to test the validity of long established models without using sample survey data and avoiding (for the most part) issues of sampling and non-sampling error, sensitizing respondents or limited observations.

This paper examined the beta-binomial model used to estimate individual magazine reach and frequency that was developed a half century ago and found it to be an excellent fit for digital issues. Equally relevant, we fit the model against passive, rather than diary-based data! Having validated the beta-binomial model, GfK MRI can integrate the R&F models for print and digital net reach as a single vehicle.

These findings further indicate the direction of a hybrid solution for integrating print and digital edition audience estimates and profiles (See Mattlin and Gagen presentation at this Forum). At present, the penetration of digital editions is relatively small and would require prohibitively costly, and substantial sample sizes (especially if these are probability-based samples from a complete frame) to report these audiences accurately and reliably. On the other hand, census-based passive data provide the requisite baseline estimates for average-issue audience and reach and frequency. The possibility of fusing large on-line samples to capture reading behavior, magazine duplication, issue-to-issue turnover and consumer profiles with passive data benchmarks seems viable. Validating the R&F model in this study is a step in that direction.

[5] All estimates are based on standardized ratings; they do not reflect audience sizes for the U.S. population
[6] We restricted the number of comparisons to 11 simply for space considerations.
[7] As stated before, some of the companies could not provide the passive data for all surveyed magazines

References

Agostini, J.M. How to Estimate Unduplicated Audiences. *Journal of Advertising Research*, Vol 11, No 3, pp 11-14
Ephron, Erwin Caution: Dangerous Curves Ahead *Ephron on Media*, March 1, 1993
Hyett, G.P. The Measurement of Readership Statistics Seminar, London School of Economics, February 1958
Smit, Edith G., Neuens, Peter C. The March to Reliable Metrics: A Half-centur of Coming Closer to the Truth *Journal of Advertising Research*, Vol 51, No.1 Supplement, pp. 124 - 135

Appendix A: Comparisons between Passive & Modeled data reach curves

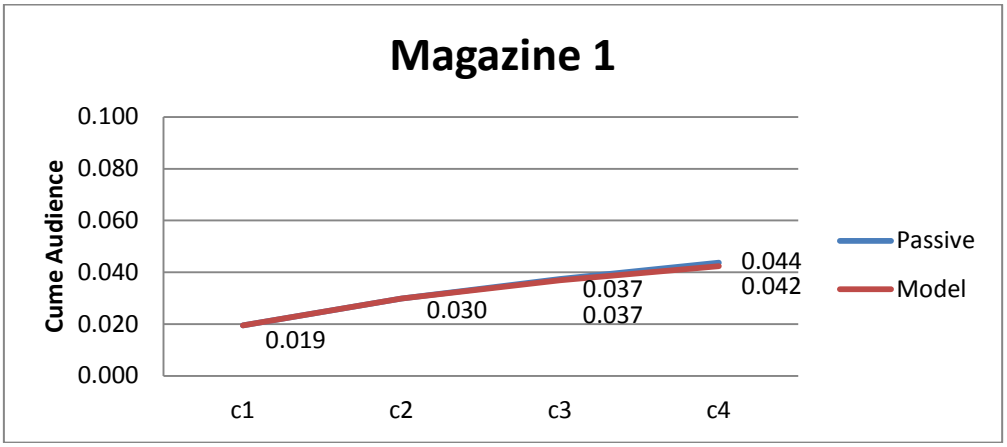

Figure 1

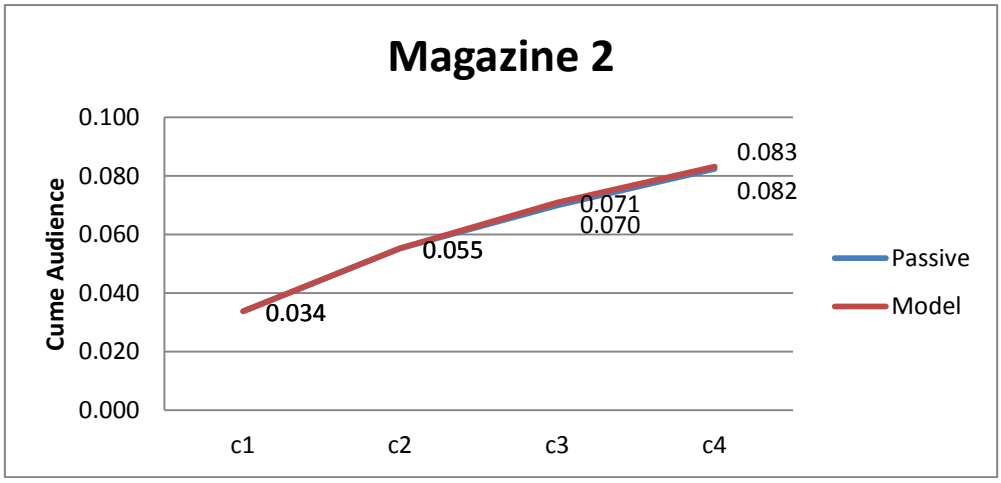

Figure 2

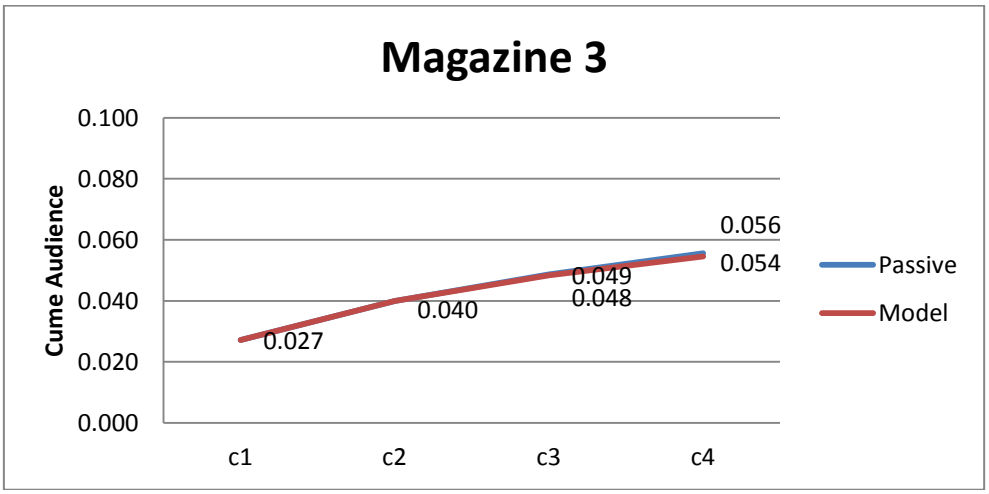

Figure 3

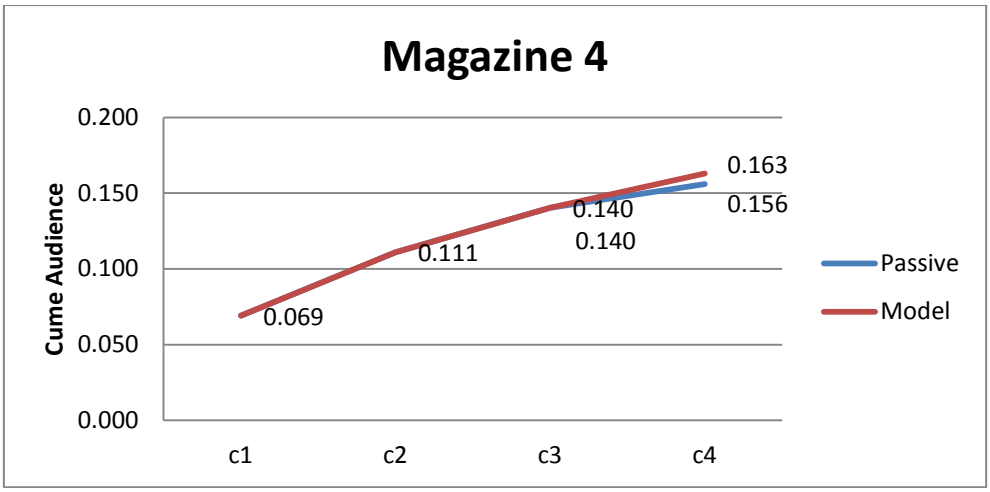

Figure 4

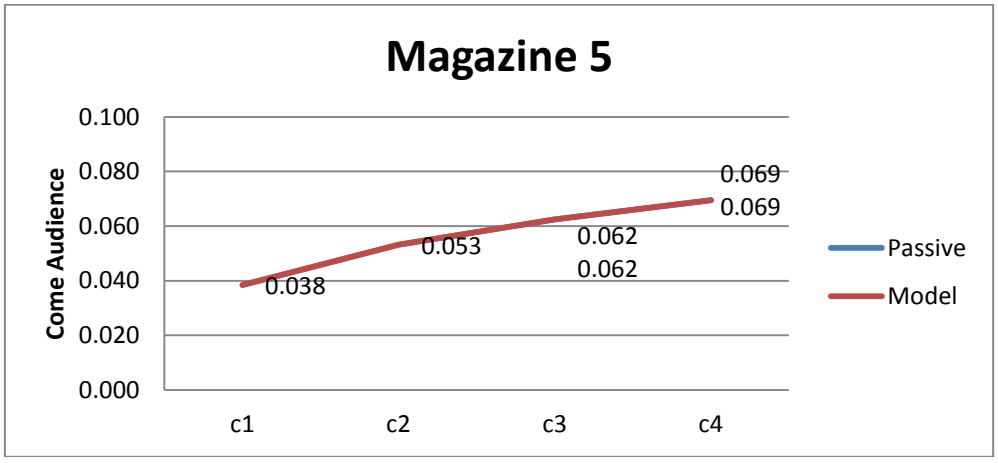

Figure 5

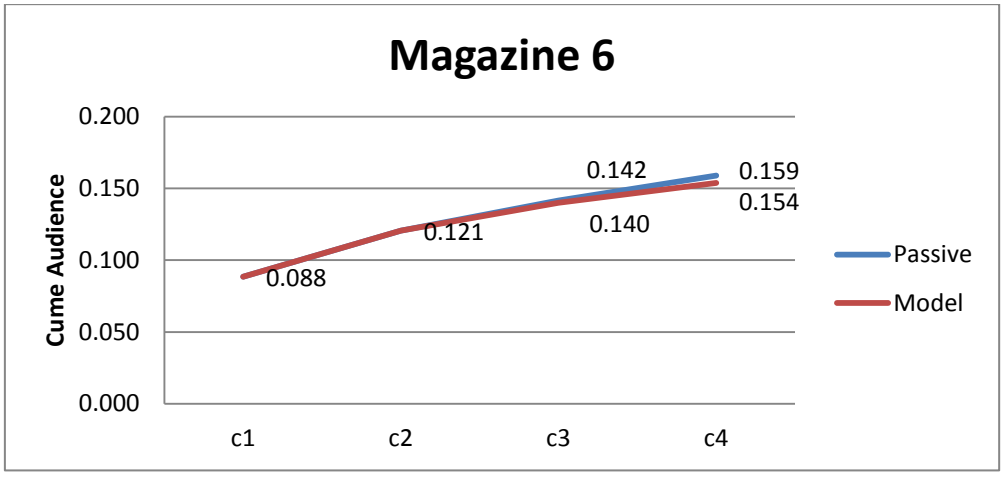

Figure 6

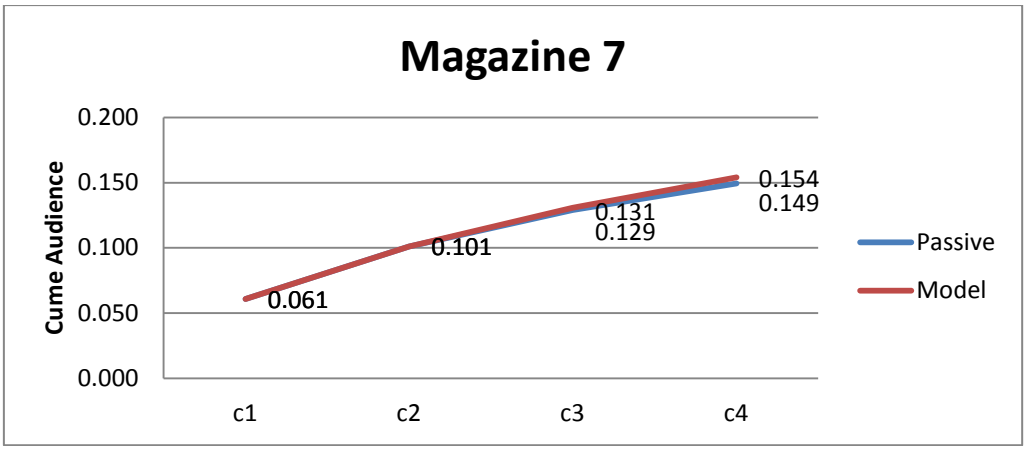

Figure 7

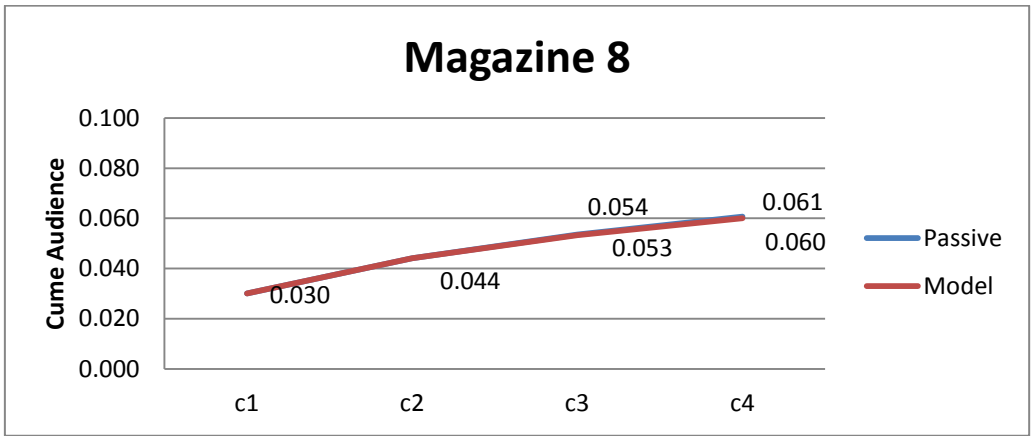

Figure 8

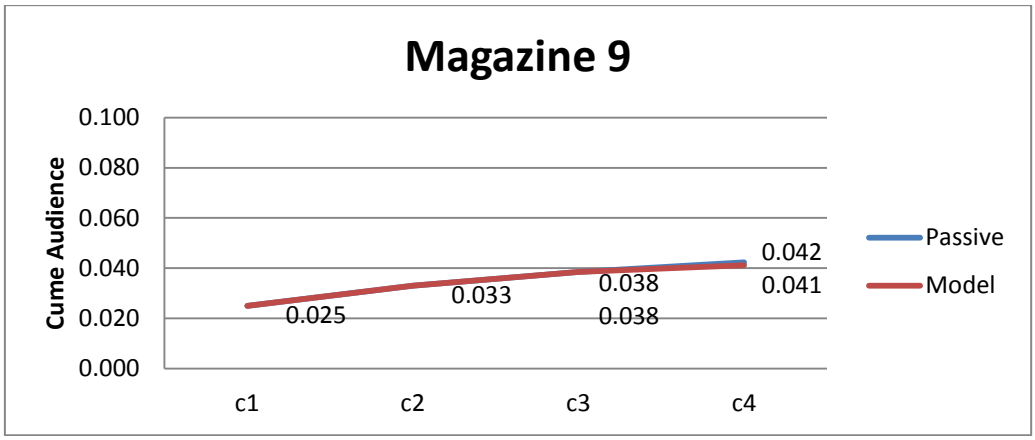

Figure 9

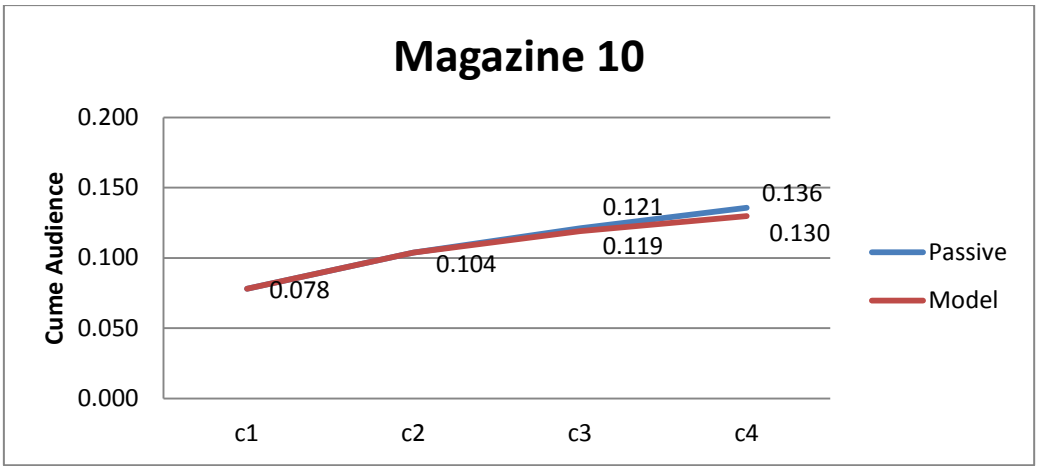

Figure 10

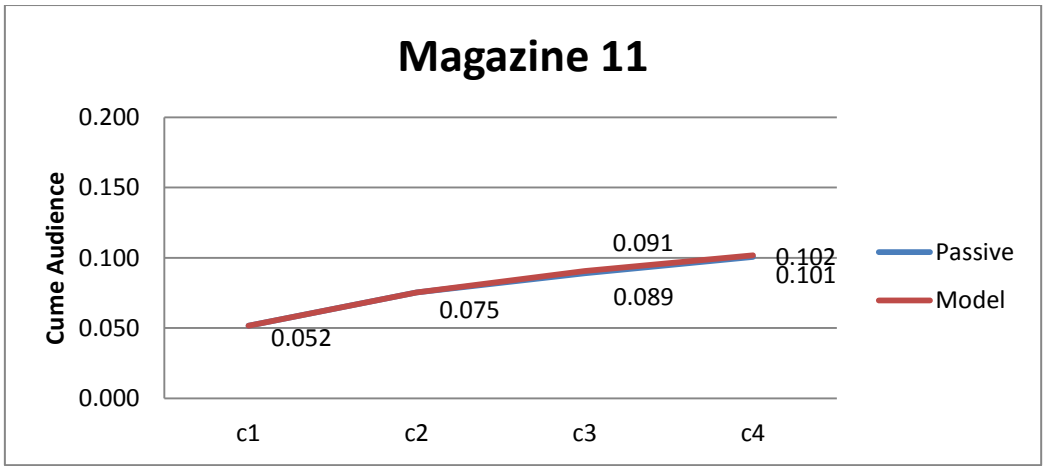

Figure 11

Appendix B: Comparisons between Passive & Modeled data reading frequency out of 4-issue reach

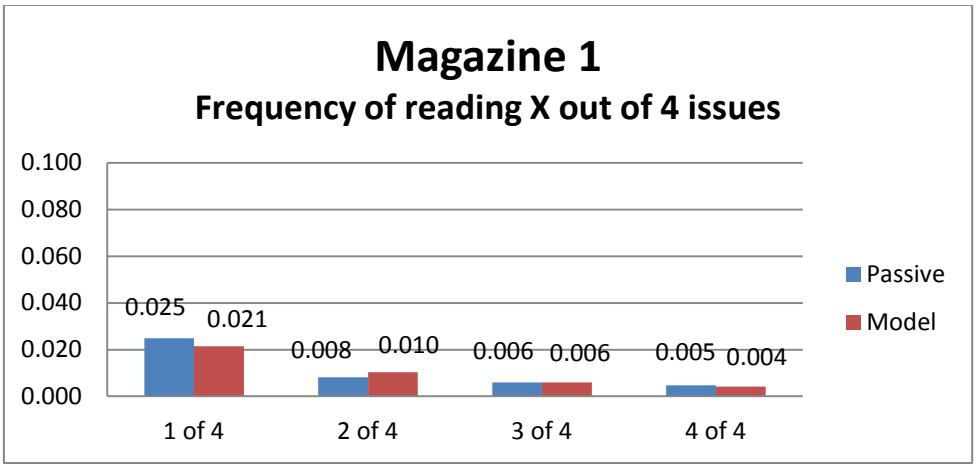

Figure 12

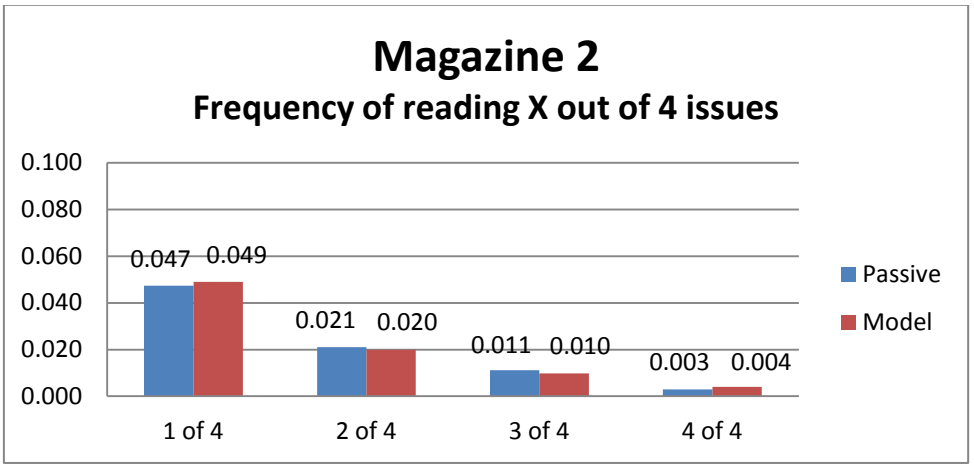

Figure 13

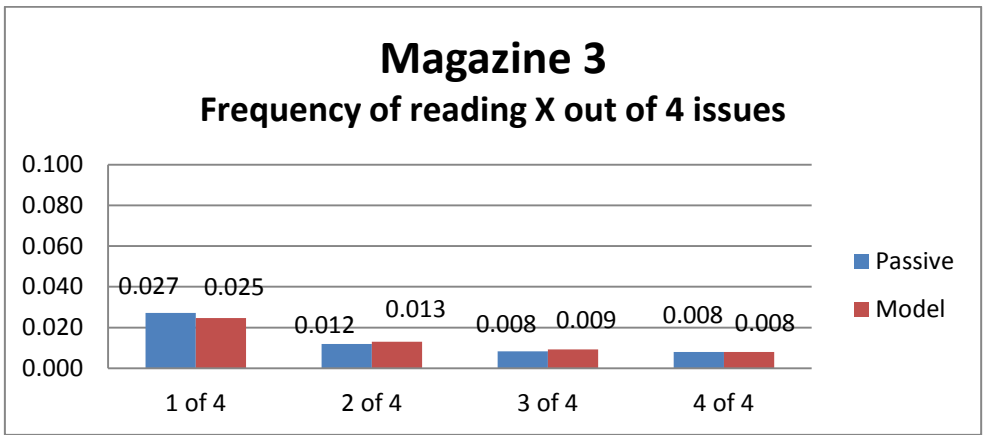

Figure 14

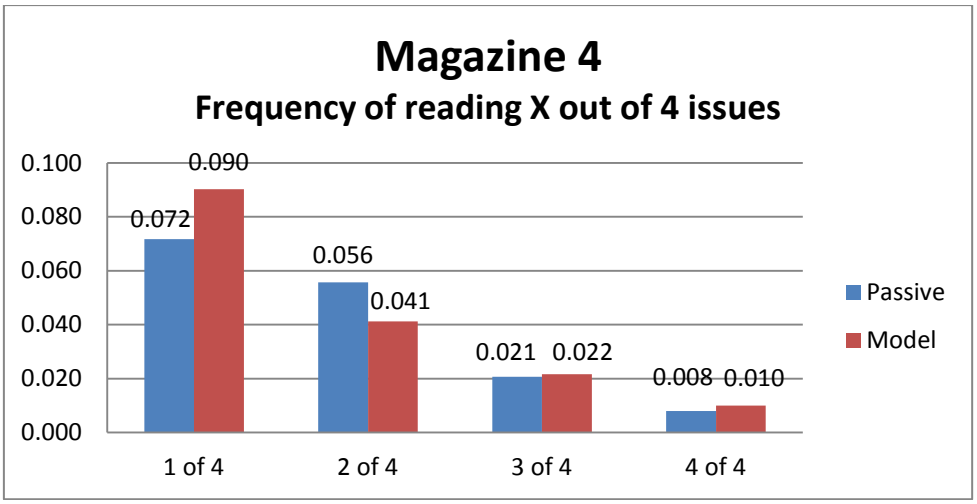

Figure 15

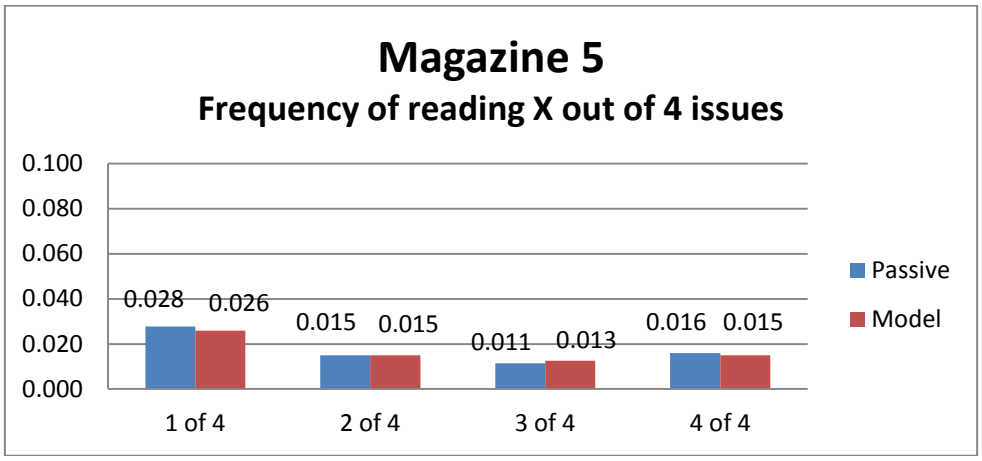

Figure 16

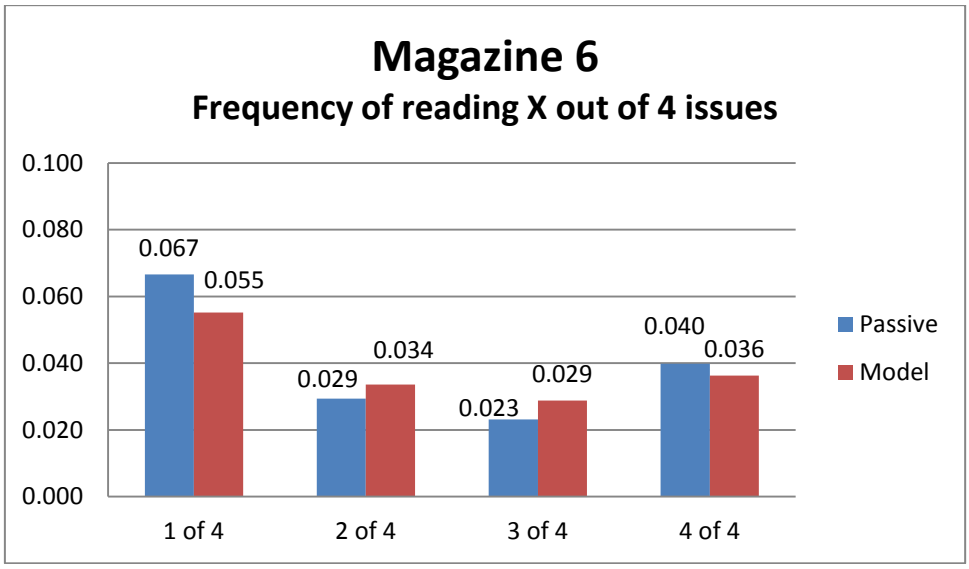

Figure 17

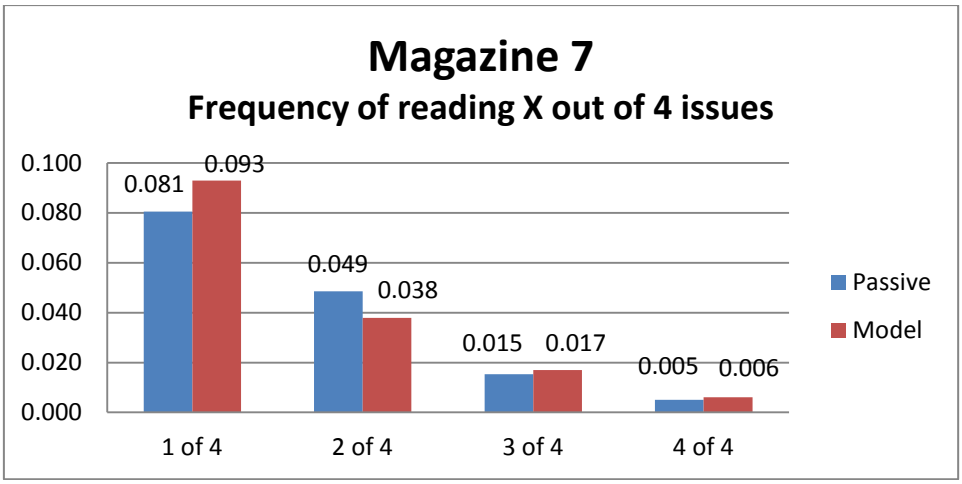

Figure 18

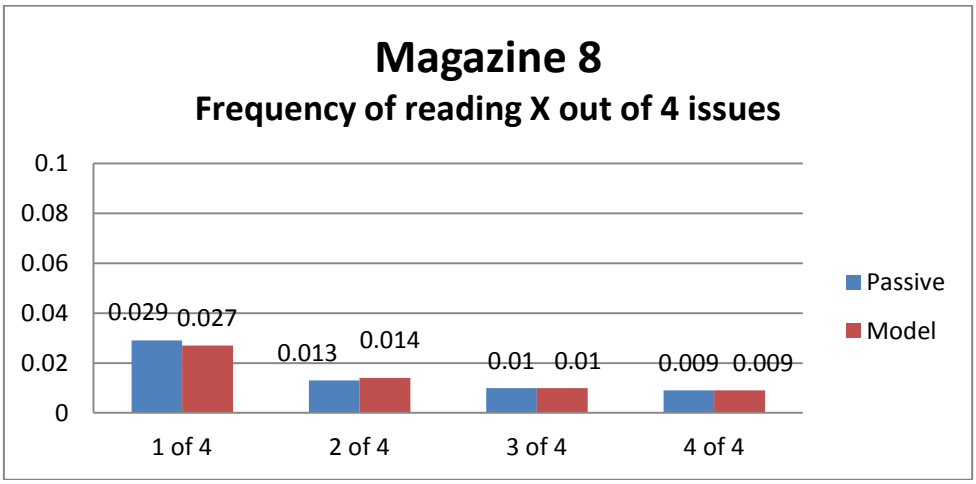

Figure 19

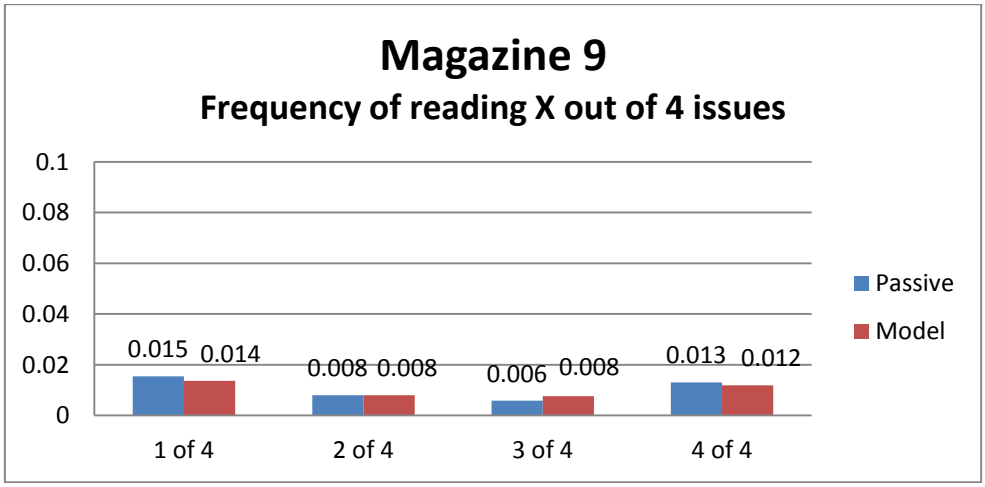

Figure 20

Appendix C: Comparisons between Survey & Modeled data reach curves

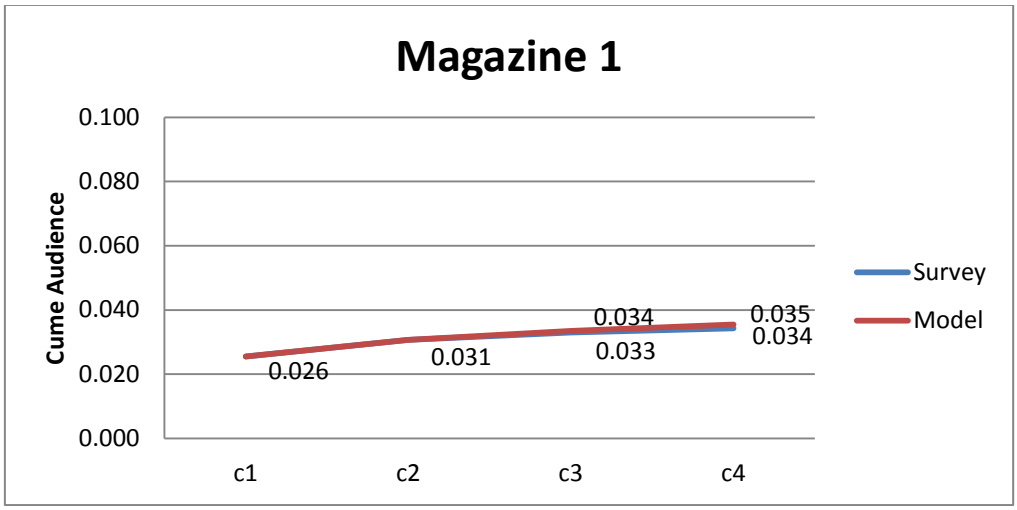

Figure 21

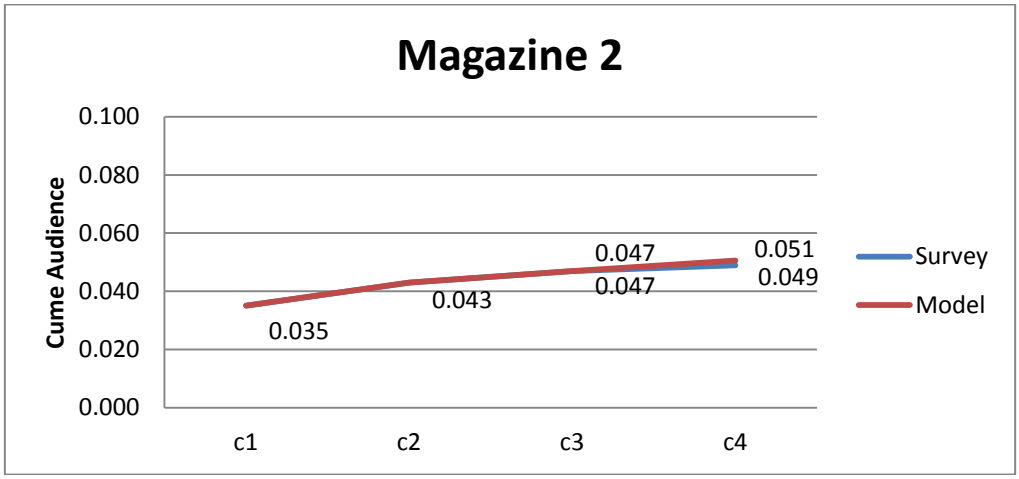

Figure 22

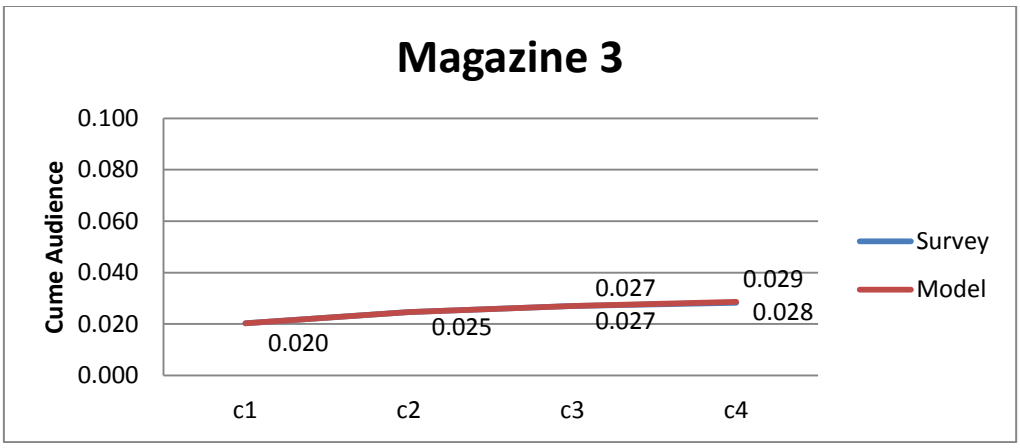

Figure 23

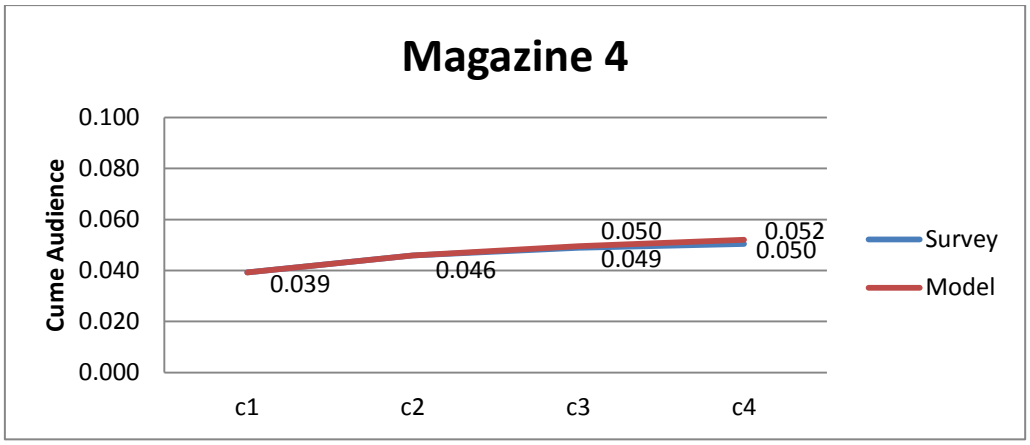

Figure 24

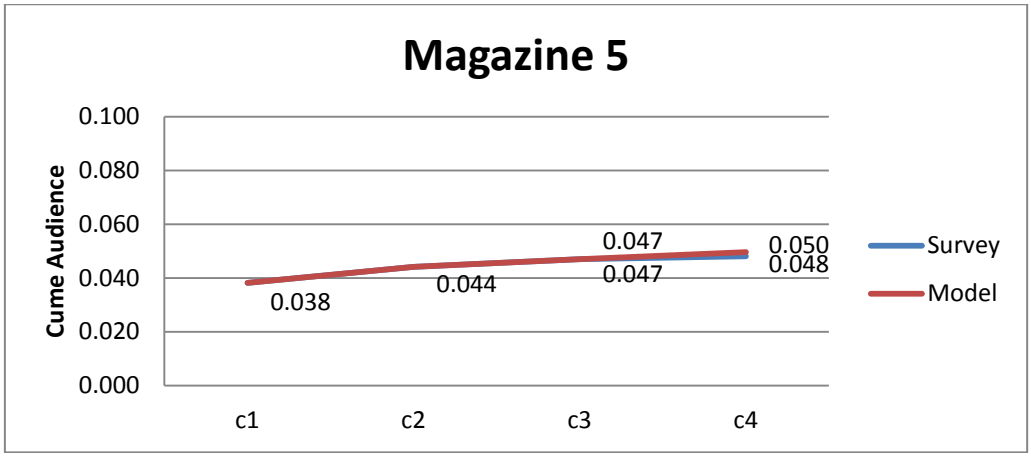

Figure 25

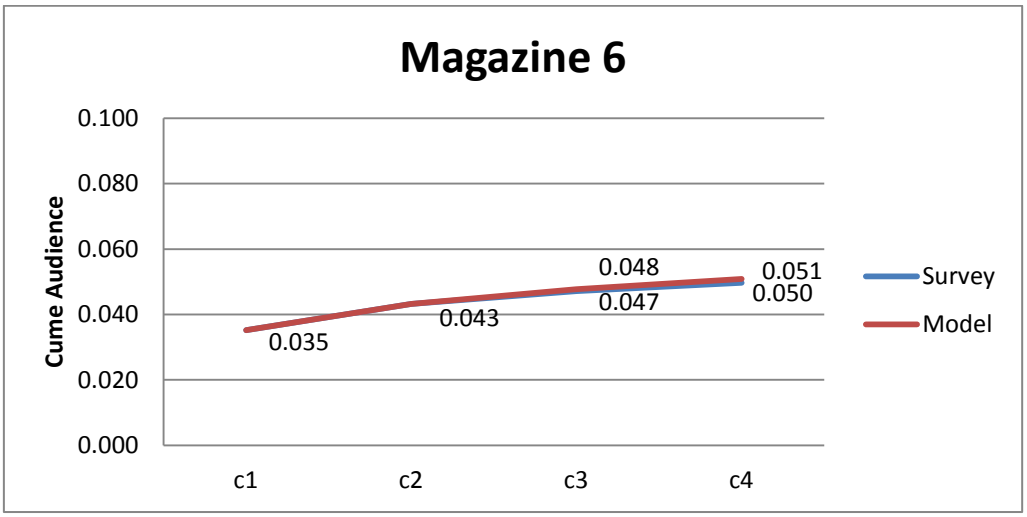

Figure 26

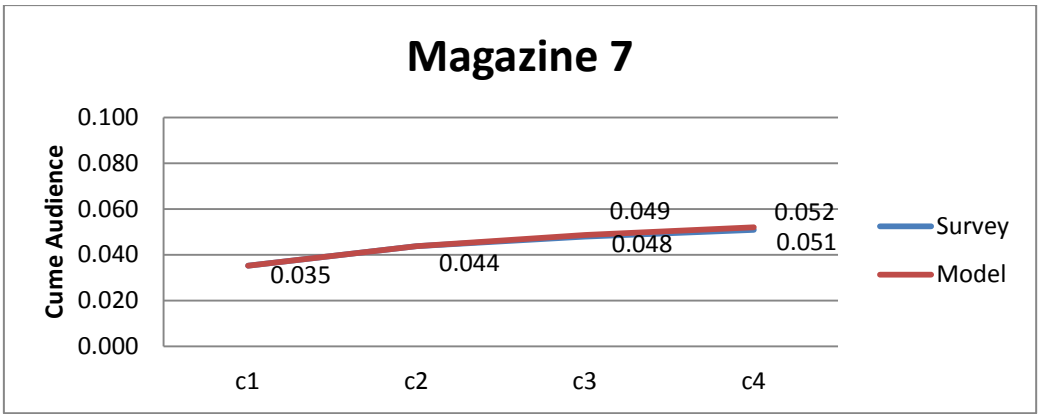

Figure 27

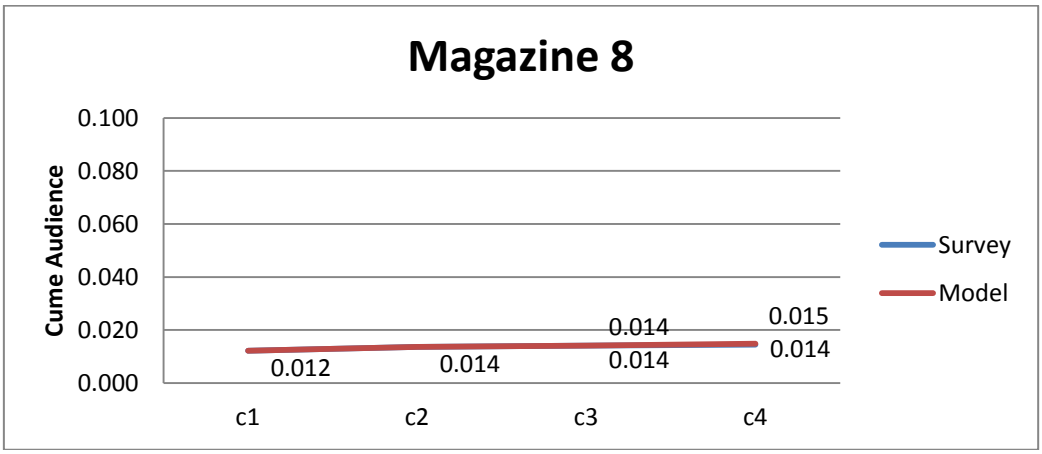

Table 28

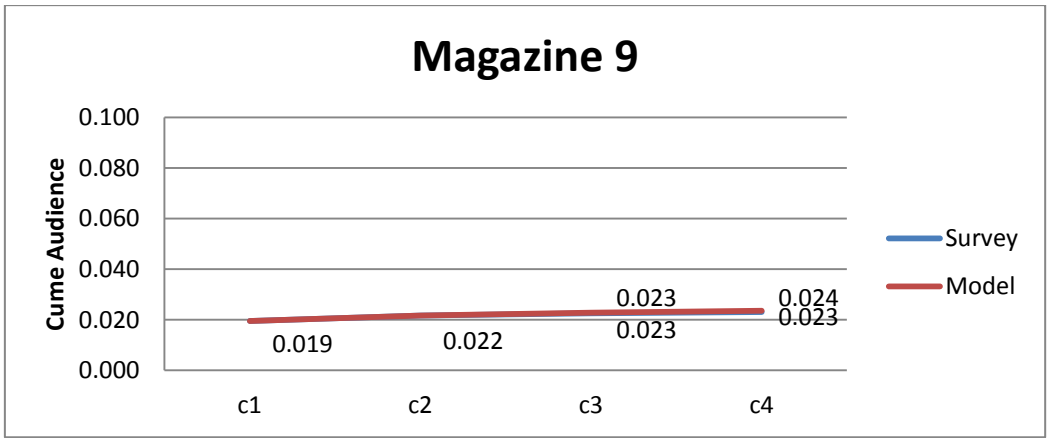

Figure 29

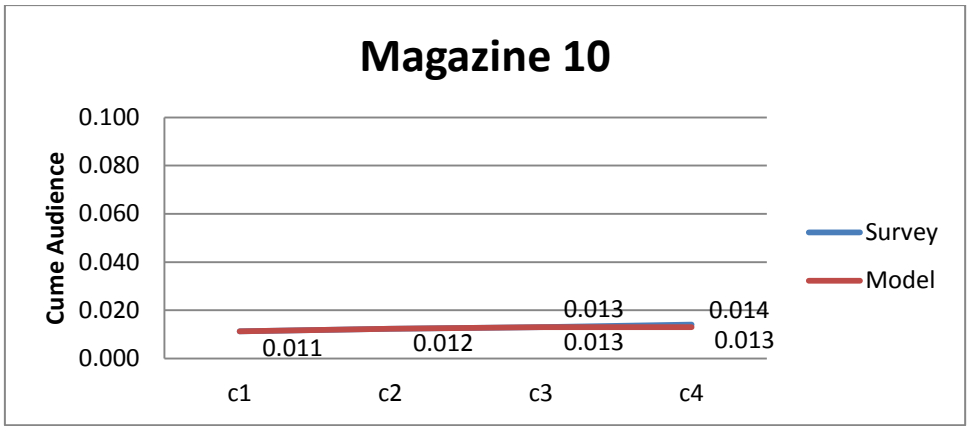

Figure 30

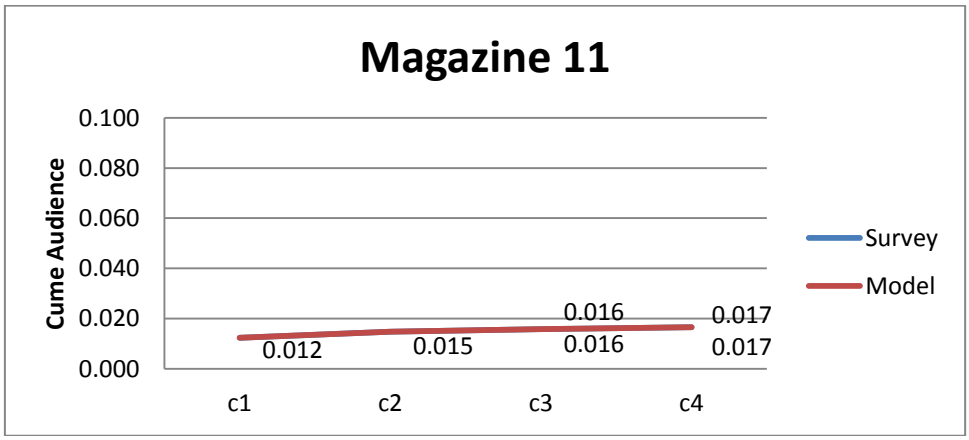

Figure 31