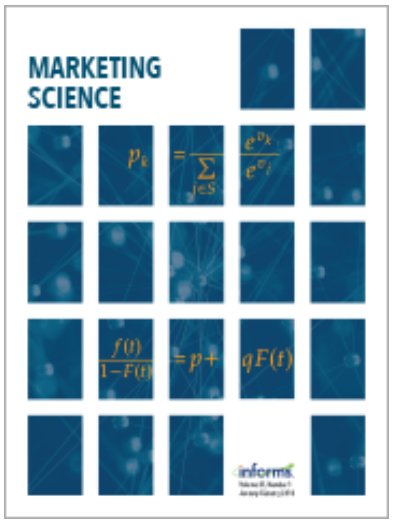


Marketing Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Practice Prize Winner—Pricing Digital Content Product Lines: A Model and Application for the National Academies Press

P. K. Kannan, Barbara Kline Pope, Sanjay Jain,

To cite this article:
P. K. Kannan, Barbara Kline Pope, Sanjay Jain, (2009) Practice Prize Winner—Pricing Digital Content Product Lines: A Model and Application for the National Academies Press. Marketing Science 28(4):620-636. https://doi.org/10.1287/mksc.1080.0481

MARKETING SCIENCE
Vol. 28, No. 4, July–August 2009, pp. 620–636
ISSN 0732-2399 | EISSN 1526-548X | 09 | 2804 | 0620

informs

DOI 10.1287/mksc.1080.0481

Practice Prize Winner

Pricing Digital Content Product Lines: A Model and Application for the National Academies Press

P. K. Kannan
Robert H. Smith School of Business, University of Maryland, College Park, Maryland 20742,
pkannan@rhsmith.umd.edu

Barbara Kline Pope
The National Academies, Washington, DC 20001, bkline@nas.edu

Sanjay Jain
Mays School of Business, Texas A&M University, College Station, Texas 77843,
sjain@mays.tamu.edu

We examine the problem of how a content provider, specifically the National Academies Press (NAP), can optimally price the different forms of its product—print and PDF—that it sells online. Whereas products in the traditional product line generally tend to be substitutes, the different content product forms could range from being substitutes to being complements across customers. Thus the content provider can possibly sell bundles of the product forms, leading to additional revenue. We first discuss NAP's decision context and describe the model we proposed for developing NAP's optimal pricing policies for its different forms. We describe the choice experiment we conducted on the publisher's website that maximally uses the online interface to collect relevant data needed to estimate our model. We show how NAP embraced the results from the model for developing a new business model and how it used the insights derived from the study to set pricing policies and monitor sales performance as a function of pricing. Finally, we perform validation of the model and the implemented policies using dynamic modeling of sales data from NAP's website. The paper illustrates how e-commerce technologies can lead to the development of optimal policies using marketing models.

Key words: pricing; product line; digital products; product form; bundling; choice experiment; Internet
History: Received: January 30, 2008; accepted: December 1, 2008; processed by Gary Lilien. Published online in *Articles in Advance* May 19, 2009.

1. Introduction

The Internet channel has transformed the marketing and distribution of both traditional printed content and multimedia content. This transformation has challenged publishers and content providers of all types who are faced with issues of optimally pricing the different products that make up their product line for online markets. The specific focus of this paper is on the National Academies Press (NAP), the National Academies'[1] book publisher that has the capability to sell and distribute books online in both print and various digital formats—ranging from a low-resolution fax-quality digital form that can be browsed at the website to a high-resolution PDF format that can be downloaded over the Internet.

Whereas this research issue may seem, at the outset, to be a standard product line design and pricing problem (Green and Krieger 1985, Dobson and Kalish 1988), there are some interesting twists. Products in the traditional product line are generally viewed as substitutes, whereas digital products can range from being perceived by customers as substitutes, to imperfect substitutes, to complements. The content (content utility) that is available is the same across the different product forms making up the product line, while the form utility that varies across forms contributes to the perception of substitutability or complementarity. Also, digital content allows for the provision of samples (or full content of a book, as in the case of our focal publisher), leading to significant and unique impacts on the demand for different products in the product line. Although extensive free content allows consumers to determine with certainty the fit of the content to their needs, it could lead to customers substituting the free content for the full product, thereby cannibalizing sales. On the other hand, for some customers free online samples might act as complements, increasing their willingness to pay for the full product after they have used a sample.

The underlying issue facing NAP in 2002 was determining how customers perceived different content

[1] National Academy of Sciences, National Academy of Engineering, Institute of Medicine, and National Research Council.

formats on a continuum from substitutes to complements and answering the following specific questions as it contemplated rolling out the PDF formats of its books for the first time:

- How should NAP price the various product forms?
- Should it offer a bundle of the product forms at a discounted price?
- What would be the impact of its policies on revenues, profits, and content penetration?

The product line pricing policy that NAP adopted was critical because, given its home within the National Academies, NAP has an extra mission beyond the financial goal of being self-sustaining; it also has to strive to disseminate knowledge from the books and reports it publishes as far and wide as possible.

In this paper, we focus on NAP's pricing problem in a monopolist context and develop an empirical model and methodology to develop pricing decisions for the online product line. We also set pricing policies for NAP while accounting for some of the complexities of the problem highlighted above. Specifically, we model the purchase of products in the product line under substitute or complement situations, thus accounting for customers' heterogeneity on that dimension. We estimate customers' preference parameters for the various product forms and their price sensitivity based on data collected as part of a choice experiment conducted at the publisher's website. We show how the model and the methodology were instrumental in setting sound pricing polices, and validate the implemented model recommendations using extensive sales and marketing mix data collected before and after the implementation of the policies.

Our research occupies a unique niche because it (1) centers on a problem situation where products in the product lines could be both substitutes and complements, and firms could offer special bundle pricing; and (2) proposes a model to account for customers' substitutability or complementarity parameters, and estimates the parameters of the models using the special features available at the publisher's website. The contribution of this paper is threefold. First, we show the existence of heterogeneity in perceptions of substitutability or complementarity of content forms among customers, and the importance of accounting for this in developing pricing policies for a wide range of emerging digital forms, a situation that is applicable to a variety of content providers. Second, we provide a compelling illustration of how Internet technologies can be used in online retail contexts to design innovative experiments to better measure and understand customer behavior, and to provide a methodology to derive optimal pricing policies and validate the results. Finally, we show how NAP management, under pressure from its various stakeholders, used the results from the marketing study to develop a new and viable business model to meet its dual mission. The success of the application of the marketing science model is validated further by using persistence modeling techniques on actual sales data to tease out the impact of the implemented policies.

The rest of this paper is organized in broad sections. In §2, we describe the product market within which NAP operates and explain the pricing dilemma the publisher faced in 2002. In §3, we present the model and methodology to optimally price the products in the product line, along with the choice experiment. Section 4 provides details of the data collected and the estimation and analysis. Section 5 describes the application of the normative and empirical findings in the development of optimal prices for the product line, and in setting up pricing policies and their implementation. In §6, we use state-of-the-art time-series modeling of the sales data to validate our model recommendations. We conclude in §7 by summarizing in broad terms the contributions of our research—specifically, the transportability of our model to other settings—and by outlining the limitations of our work.

2. Research Background

2.1. NAP Background

The product market of the National Academies Press consists of scholarly, research-oriented publications, a market dominated by university presses and other scholarly publishers.[2] Given its dual mission of dissemination and self-sustainability, the nonprofit NAP is motivated to build profit and recapitalize their program to build new technologies that enhance their reach to the audiences for their content. When NAP obtains content to be published from its institutional authors, it is usually specialized, unique, and authoritative. The findings of the research studies are available only from the National Academies. NAP does not compete with other publishers for the right to publish the content; they have the exclusive right to publish all studies of the National Academies. Thus, given the specialized and authoritative nature of the content, and lack of direct competition, we can consider the market for content at the customer level to be monopolistic.[3]

2.2. Context of the Decision

NAP took a leadership role in dissemination when, in 1994, along with selling printed books online, they

[2] NAP publishes about 200 titles a year—mainly scholarly monographs, study reports, and reference material in all areas of science, education, engineering, and health and medicine, which are specialized and focused, catering to several distinct niche markets such as researchers and practitioners with well-defined needs.

[3] This is quite typical of the $3.3 billion scholarly publishing market where publishers can be considered local monopolies because they own the rights to the intellectual content they publish.

started offering the full text of the books they sell in page-by-page fax-quality format for free to anyone in the world through its website. However, in 2001, the early demand for PDF formats, and the ease with which the PDF format could be produced and distributed online, made NAP contemplate selling high-quality PDF files in addition to the printed form. However, NAP did not know what proportion of the customer base would purchase PDF formats or how much they would pay for them. More important, it did not know how PDF offerings would affect the demand for the printed format, and in turn, print-based revenues.

2.3. Stakeholder Pressure

The scientists leading the National Academies, and those who produced the content that NAP sold, were interested in using the Web to disseminate scientific knowledge around the world. They were pushing NAP management very hard in early 2002 to provide the complete full text of their studies in PDF format free of charge to everyone on the Web. The leadership believed that the two forms—PDF and print—would be strong complements, and thus NAP's revenues would not be affected. However, if the two forms were actually substitutes, providing free PDFs could be disastrous to the NAP's bottom line. For NAP management this was an issue of critical importance. NAP's self-sustainability and the profits necessary for investing in projects that enhance customers' experience at their website were at stake. Thus they embarked on a marketing science study that led to this research, which was supported by a grant from the Andrew W. Mellon Foundation.

2.4. Extant Research

In a traditional product line pricing situation, the products are generally viewed by customers as substitutes, because in many common product categories such as consumer durables and consumer package goods, the usage situations tend to be similar. Thus, except for some notable exceptions (Reibstein and Gatignon 1984), extant research has framed the problem treating the products in the product line as substitutes (for example, Mussa and Rosen 1978; Dobson and Kalish 1988, 1993; Kohli and Sukumar 1990; Moorthy and Png 1992; Cohen et al. 1997). However, there are many situations where digital products can be viewed as imperfect substitutes or even complements. For example, a customer buying a book in electronic form could also be interested in purchasing a printed version of the book because the usage situations of these forms, and the benefits they provide, can be different and therefore complementary. On the other hand, for some customers these product forms can be perfect substitutes. Some people covet the convenience of a printed and bound book, while others find that a PDF version of a book meets their needs perfectly. Thus, one could view the universe of customers to be heterogeneous in terms of their preference parameters for the product forms that render these product forms perfect substitutes, or independent, or complements.

Our work is also related to the literature on bundle pricing (see, for example, Adams and Yellen 1976, Farquhar and Rao 1976, Schmalensee 1984, Venkatesh and Mahajan 1993, Chung and Rao 2003) that includes studies of situations that allow customers to buy multiple product forms or components. Whereas Venkatesh and Mahajan (1993) examine pricing in the pure bundling and mixed bundling contexts for performances, Chung and Rao (2003) determine pricing for the pure bundling case for components using state-of-the-art Bayesian estimation techniques. In addition, Venkatesh and Chatterjee (2006) have examined pricing of magazines and unbundling content in the online context. However, extant approaches have not allowed for the possibility that products could be complements for some consumers and substitutes for others. Specifically, Venkatesh and Kamakura (2003) argue that customers are heterogeneous in their substitutability and complementarity perceptions, and this heterogeneity in the so-called "degree of contingency" plays an important role in the bundling and pricing strategies. It is precisely this parameter we account for and estimate in our model, albeit at the product form level instead of using conjoint analysis (cf. Kohli and Mahajan 1991).

2.5. Free Browsing

Free browsing on NAP's website is the online correlate to someone flipping through an entire book at a bricks-and-mortar Barnes & Noble bookstore. It is useful to our research because it ensures that there can be no uncertainty about the content of the title that a customer considers and buys. Free browsing could stimulate the demand for other product forms in the product line (print or PDF), and it could also be a potential substitute. Because NAP considered free browsing of the entire content as a permanent feature, it became the baseline for estimating the marginal utilities of the various forms, as well as attenuated the impact of content piracy on our estimates of profits. (For analysis of the impact of piracy on profits, see Sinha and Mandel 2008, Jain 2008).

Finally, although NAP is a nonprofit small business, it is motivated to seek a profit maximization objective so that it can use the profits to capitalize its programs and offer superior customer browsing experiences through technological features. Thus, while a traditional nonprofit firm's objective could be to maximize sales or penetration (see, for example, Liu and

Weinberg 2004), NAP chose to maximize profits given a certain level of penetration in seeking answers to its questions.

3. The Model and Methodology

The utility that a consumer derives from purchasing either the PDF form of the book or the printed form depends on two factors. First, the utility depends on the degree of fit between the contents of the book and the individual consumer's needs. This degree of fit is common across the two forms, as the content that NAP publishes in each form is identical. Second, each consumer may value the product forms differently. Thus, some consumers prefer the printed version to the PDF file and some others might value the PDF form higher than print. The utility that a consumer i derives from purchasing a title (content) in product form j is given by

$$U_{ij} = \beta_{ij} X_i - \beta_{pi} p_j + \varepsilon_{ij} \quad j = \{1 = \text{print}, 2 = \text{PDF}\},^4 \quad (1)$$

where X_i is individual i's measure of the degree of fit of the content to his or her needs, β_{ij} is a parameter that measures the value that customer i places on product form j, β_{pi} is the price sensitivity for customer i, p_j is the price of product form j, and ε_{ij} is the error term that we assume is double-exponentially distributed. The consumer also has the option of not buying either product form, which is assumed to provide a zero (marginal) utility to the consumer. (The notion of zero utility is relative to the free browsing of content, to which a customer has full access at all times.)

Finally, the consumer also has the option of buying both product forms. This is relevant only when the consumer does not perceive the product formats to be perfect substitutes. In the case of books, the PDF form and the printed form can often be viewed as serving complementary usages, and consumers may be interested in purchasing both (as we discussed above). We model this aspect by assuming that the utility (U) that a customer derives from purchasing the bundle b is given by

$$U_{ib} = (\beta_{i1} + \beta_{i2} + \Delta\beta_i) X_i - \beta_{pi} p_b + \varepsilon_{ib}, \quad (2)$$

where $\Delta\beta_i$ is a parameter that represents how consumer i values the bundle in a marginal sense, and p_b is the price of the bundle where $p_b \geq \min(p_1, p_2)$. If $\Delta\beta_i = -\min(\beta_{i1}, \beta_{i2})$, then the customer perceives the two product forms to be perfect substitutes and will only purchase one of the two product forms. However, if $\Delta\beta_i > -\min(\beta_{i1}, \beta_{i2})$, then the consumer derives some additional utility in buying the second form, given that he has already decided to buy one form of the product. For example, a consumer may find the PDF form provides some additional utility even when he already has purchased the printed book. Note that if $\Delta\beta_i > 0$, then the utility that the customer obtains from buying the bundle exceeds the total utility from the two product forms separately. In other words, the two product forms are strong complements. For intermediate values, $-\min(\beta_{i1}, \beta_{i2}) < \Delta\beta_i < 0$, the product forms act as partial substitutes for the customer. Thus, $\Delta\beta_i$ captures the form-preference interdependencies at the individual consumer level that determine their utilities for the bundle of forms.

Given customers' choice and degree of fit data, the parameters of the above model are estimated using a finite mixture (FM) model,[5] which captures the heterogeneity at the consumer level with discrete segment-level estimates for preference parameters (Kamakura and Russell 1989).

3.1. The Optimization Problem

We assume that the distribution function of X_i across the population is given by $F(X)$ with support Θ and where $F(\cdot)$ could be a continuous distribution or a discrete one. Given that we estimate discrete segments of consumers with parameters varying across segments using a finite mixture model, assume that there are K such latent segments. Let the parameters in the kth segment be denoted by $(\beta_1^k, \beta_2^k, \beta_b^k, \beta_p^k)$. Given the assumption that the error terms are double-exponential distributed, the expected market share in segment k for the print form is given by

$$E(MS_1^k) = \int_\Theta \left[\frac{\exp(U_1^k(X))}{1 + \exp(U_1^k(X)) + \exp(U_2^k(X)) + \exp(U_b^k(X))} \right] dF(X), \quad (3)$$

$$E(MS_1^k) = \int_\Theta [(\exp(\beta_1^k X - \beta_p^k p_1)) \cdot (1 + \exp(\beta_1^k X - \beta_p^k p_1) + \exp(\beta_2^k X - \beta_p^k p_2) + \exp((\beta_1^k + \beta_2^k + \Delta\beta^k) X - \beta_p^k p_b))^{-1}] dF(X). \quad (4)$$

If the marginal cost of the print form of a title is c_1, and the marginal cost of the PDF form c_2, then the

[4] Equation (1) is conceptually similar to Item Response Theory (IRT) models or latent trait models (Lazarsfeld and Henry 1968, Dayton 1998). In an IRT model, individual i's degree of fit (the term X_i in our model) is replaced by a latent individual-level discrimination or fit parameter θ_i, which are estimated using self-reported fits. We choose to use the self-reported degree of fit directly in the model as in Equation (1) because it performs better in out-of-sample predictions as compared to the case when we estimate fit parameters. This is discussed later in the paper.

[5] Alternatively, one can also use hierarchical Bayesian estimation methods.

expected profits from the print form of the title—when the print form is the only form purchased—can be given by

$$\Pi_1 = (p_1 - c_1) \sum_{k=1}^{K} \delta^k E(MS_1^k), \tag{5}$$

where δ^k is the estimated size of the latent segment k.

Similarly, the expected profits from the PDF form of the title, when PDF is the only form purchased is given by

$$\Pi_2 = (p_2 - c_2) \sum_{k=1}^{K} \delta^k E(MS_2^k). \tag{6}$$

When the two forms are purchased together as a bundle, then the expected profits are given by

$$\Pi_b = (p_b - c_1 - c_2) \sum_{k=1}^{K} \delta^k E(MS_b^k). \tag{7}$$

The total profit from the title is given by

$$\Pi = \Pi_1 + \Pi_2 + \Pi_b. \tag{8}$$

A content provider's problem is to choose p_1^*, p_2^* and the bundle price p_b^* to maximize total profits. In our optimization, we will first consider the case when there are no discounts for buying the bundle, that is, when $p_b = p_1 + p_2$. The optimization problem then becomes

$$\begin{aligned} &\max_{p_1, p_2, p_b} \ \Pi(p_1, p_2, p_b) \\ &\text{subject to } \ p_b = p_1 + p_2. \end{aligned} \tag{9}$$

If we allow the possibility that consumers can get a discount for purchasing the bundle, then the optimization problem becomes

$$\begin{aligned} &\max_{p_1, p_2, p_b} \ \Pi(p_1, p_2, p_b) \\ &\text{subject to } \ p_b \le p_1 + p_2. \end{aligned} \tag{10}$$

We use an adaptive search algorithm along with fine-grid search methods to determine optimal prices. In some scenarios under consideration, we constrain the total market penetration in the optimal solution to be at least some specific level (existing levels, for example). These constraints are taken into account while performing grid search to reject solutions outside the feasible domain. The specifics of obtaining the estimates that go into the optimization, the distribution of $F(X)$, and the data collection procedures are discussed in the following subsection.

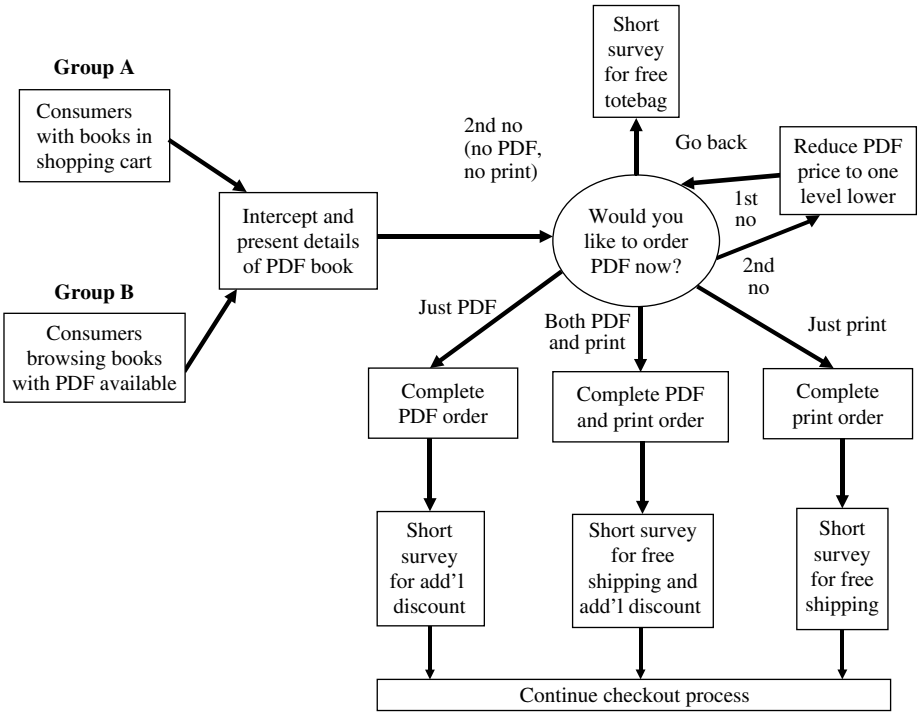
3.2. The Online Choice Experiment

The choice experiment included about 500 titles that the publisher was already selling in printed format at their website for at least two months prior to the experiment. The prices of these books in printed form were already fixed and thus were probably common knowledge to customers frequenting the website. Hence, the prices for the printed books were kept at the existing level throughout the experiment to minimize confounds from reference effects. The content of these titles was also available for free browsing in its entirety at the website. The PDF versions of the 500 titles were being made available for the first time during the experiment at the website, and the information about the PDF format was provided only to the participants in the experiment as per the design of the experiment. The choice experiment was also designed so that there was minimal disruption to the existing transactions at the publisher's website.

To estimate our choice model, we needed to obtain choice data from those visitors to the NAP website who had some interest in titles sold by NAP and would consider purchasing a title—namely, the group of potential buyers. However, even though the number of visitors to the website was large, 65%–70% of the visitors quit after viewing the site landing page. Additionally, visitors' interests were spread across many subjects and titles, with an overall purchase conversion rate of less than 0.6%. Given these conversion rates, randomly intercepting any visitor at the website would not be very useful because of the choices necessary in the experiment (see Bucklin and Sismeiro 2003, Moe and Fader 2004). Choosing or at least showing interest in a title was a necessary condition to define the population of interest. Thus, to be efficient and effective, our sampling scheme had to focus on those customers who showed some interest in one of the 500 titles included in the experiment rather than target visitors randomly at the site. Therefore, customers were grouped into one of two categories for designation as participants in the experiment (Groups A and B) as shown in Figure 1. Each group was defined on the basis of the location in the website where they were targeted.

Those in Group A were the customers who showed their intent to purchase, evidenced by having a printed book in their shopping cart and having clicked on the checkout button. After clicking on the checkout button, they were randomly intercepted (one in two) and presented with the details of the PDF format of the title they had in their shopping cart. The details of the PDF format included information about the quality of the PDF, a clickable button to view the sample PDF, download time, and the price of the PDF format in U.S. dollars. Free browsing was well advertised and prominent at NAP's website, so we assumed that these customers knew they could peruse the content to compare the fit of the content to their needs. They most likely took advantage of this feature and browsed the

Figure 1 The Choice Experiment



content online, having prompted their visit by receiving an e-mail or other promotion material, reading a news article or a review, or otherwise being informed about the content. Customers in Group B were those who were browsing the book title for which a PDF form existed. This group was intercepted (one in four) at random after perusing five pages of content and was presented with details of the PDF in the same fashion as those in Group A.

The specific choice sequences for customers in both groups were designed to obtain at least two choice observations from each customer for our estimation purposes.[6] For example, Group A customers were observed making a choice between no purchase and the printed book, after which they were intercepted. Then they were provided with three alternatives: (a) stick with the print, (b) switch to PDF, or (c) purchase the bundle. If they chose to stick with the print, the PDF prices were lowered one level and the options repeated (see Figure 1). Similarly, Group B customers were first offered the PDF for purchase at the specific price level. If they chose the PDF, they were provided the alternatives: (a) stick with the PDF, (b) buy the print, or (c) buy the bundle. If they chose "no purchase," the PDF prices were lowered one level and the PDF offer repeated. (The specific price levels for the print, PDF, and bundle are discussed in the next subsection.) If they picked PDF at this point, all three options—PDF, print, and bundle—were offered. These sequences of choices ensured that at least two choice observations from each respondent in both groups.

There were specific reasons why the customers were targeted in the process of purchasing printed books as well as while browsing. First, an analysis of the history of print purchases showed that, in any given day, about 40% of purchases were made directly by customers without going to the browse section. Additionally, the examination of purchase rates prior to the experiment showed that conversion rates of customers from the two locations (Groups A and B) were very close, indicating that they could be considered samples from the same population of customers with regard to purchase rates. Thus, targeting respondents at the two locations provided increased opportunities for sampling from the same population of potential buyers and increasing our sample size, an effect that proved very useful given that our experiment ran for only three weeks. Second, browsing the book indicated some level of interest in the item. Given that only the printed version of the book was available for sale, the customer might not have purchased it at the listed prices. However, with the introduction of the PDF (particularly at lower than printed book prices), a new purchasing option was offered.

[6] Given that most purchases at the NAP website tended to be single title rather than multiple, each respondent could only provide data on one title. However, to estimate our model we needed at least two choice observations for each customer on the title with which they were interacting.

Thus, in a sense, Group A customers allow us to estimate the cannibalization of print by PDF, whereas Group B customers provide us with estimates of market expansion because of the PDF form at different price levels. This scheme, however, introduces sampling biases. We control for this bias and discuss the details in a later section.

3.3. PDF Price Levels

In the choice experiment, PDF prices were set at six levels relative to the printed book prices—at 110%, 100%, 75%, 50%, 25%, and 0% of the printed book price. These specific levels were chosen to cover all the possibilities that NAP was considering with regard to the pricing strategy of PDF forms. The PDF prices were displayed in absolute dollar terms. For each title, the price of the printed form was held constant, while the PDF price was set at one of the above six levels relative to the printed book price. If a customer did not choose a PDF format at the initial price level, it was dropped one level lower, and the offer was repeated for all levels except those which were already at 0% of the print book price. This experimental scheme is very similar to contingency valuation experiments in economics (McConnell 1990, Alberini 1995), except that we do not increase the PDF price levels when the respondent makes a choice (see Figure 1).

The PDF prices for the 500 titles were randomized across three attributes: the subject category, the price level of the printed book, and the popularity of the book (represented by the number of visits to the book on the website). For example, for titles in the "education" subject category with the printed books' price level around $30, the corresponding PDFs were priced at levels ranging from 110% to 0% of the printed book price with a generally equal number of titles assigned for each of the 100% to 25% levels, and half that number of titles assigned for the 110% and 0% levels. The PDF prices were also assigned to achieve a similar treatment balance across the popularity of the books. Thus, among the $40 "most popular" books category, the PDF prices were assigned to cover all relative price ranges (110% to 0%). This ensured that the experiment results could be generalized across the entire product line. The prices of the bundles offered to visitors were the sum of the prices of the individual printed and PDF formats. There were no special bundle discounts in the experiment.[7]

[7] The lack of variation in bundle prices relative to print and PDF prices allows the estimation of form-specific price parameters for only two out of the three forms in the model; however, we did not find any significant differences in form-specific price parameter estimates in the model. Thus we used a common price parameter, β_p, for all forms in the model.

3.4. The Measure of Fit

In addition to participating in the choice experiment, each customer filled out an incentive-based survey (see Figure 1). (All incentives had similar value.) In the survey, each participant provided a response to the "extent to which the content of the book fits" his or her needs on a nine-point scale ranging from 1 (poor fit) to 9 (excellent fit). This provided the values of X_i needed for our estimation.[8] Note that measuring the fit of the book content is a more reliable measure of customers' valuation of the product than measuring the valuation directly (see Morrison 1979). The use of a nine-point scale to measure X_i ensured that the information loss in capturing the underlying continuous distribution was limited to 10% (see Srinivasan and Basu 1989).

4. Data Description, Estimation, and Analysis

The data consist of the choices made by 1,140 customers who participated in the three-week-long online choice experiment and who filled out the survey. About 10% of the data had to be eliminated because of incomplete responses to the survey or incomplete Web logs, leaving a total of 1,027 valid customer responses to the choice experiment and survey. Of this, 312 were from Group A (intercepted as they clicked the checkout button with the printed version of the book in their shopping cart) and the rest were from Group B (those browsing the page-by-page online version of the book). Overall, the data consisted of customer responses to 3,012 choice occasions. Among Group A customers, 46% ultimately purchased the printed format, 33% switched to the PDF format, and about 21% purchased the bundle. Among Group B customers, the percentage of purchases of print, PDF and bundle was 20%, 16%, and 4%, respectively, with 60% of customers purchasing nothing.

4.1. Correcting for Sample Selection Bias

A choice-based sampling scheme such as that used for Group A leads to a selection bias. This, in turn, leads

[8] We also examined an alternate specification of utility Equations (1) and (2), where X_i was replaced by a respondent-specific fit parameter θ_i similar to an IRT model (Dayton 1998). Given that each respondent provided measure of fit only on one title, we could not identify both θ_i and β_{ij} in Equations (1) and (2). Instead, we could only estimate θ_i and β_j using choice data and distribution of X_i using maximum simulated likelihood methods. In comparing the model performance with our specification of survey-based X_i, we found that our specification performed better on in-sample fit dimension and much better in out-of-sample prediction (an increase of 9% in hit rate from 49% to 58%), a critical dimension given our focus on developing pricing policies. Likewise, Zhang and Krishnamurthi (2004) also find the survey-based distance measures to be superior to the estimated ones in predictive performance.

to inconsistent estimates of the alternative specific preference parameters (Manski and Lerman 1977). While a direct weighting of the choice-based sample—as suggested by Manski and Lerman—using historical purchase rates might be useful,[9] we have an additional complexity in that the experiment leads to an increased purchase rate (as compared to historical purchase rates) because of the presence of PDF forms. The post-experiment purchase rate in Group B was 40%. Assuming that Groups A and B are random draws from the same population, we can project the same post-experiment for purchase rate (40%) for Group A. Applying the Manski and Lerman correction methodology to the post-experiment data leads to a weight of 0.497 for the Group A purchase data, a weight of 0.814 for the Group B purchase data, and a weight of 1.23 for Group B no-purchase data. These weights were incorporated in the FM model likelihood equation to ensure that the weighted exogenous maximum likelihood estimators would be consistent (Manski and Lerman 1977).

Table 1 Estimates from the Finite Mixture Logit—Proposed Model

	Estimates (standard errors)			
Parameters	Segment 1	Segment 2	Segment 3	Segment 4
β_{Print}	0.2417 (0.0519)	0.2367 (0.0343)	0.1562 (0.0292)	0.0627 (0.0137)
β_{PDF}	0.0329 (0.0062)	0.0841 (0.0198)	0.1046 (0.0217)	0.1442 (0.0309)
$\Delta\beta$	−0.0298 (0.0017)	−0.0367 (0.0087)	−0.0542 (0.0103)	−0.0253 (0.0064)
β_{Price}	−0.0712 (0.0079)	−0.0922 (0.0064)	−0.0991 (0.0048)	−0.1127 (0.0121)
λ^*	1.34117 (0.2276)	1.24286 (0.1473)	0.22437 (0.0419)	0 (Fixed)
Relative segment size (%)	40.1	36.3	13.1	10.5

	Log likelihood (LL)	AIC ($LL - p$)	BIC ($LL - p * \ln(N)/2$)
Model fit	−2,514.17	−2,533.17	−2,590.30

Notes. All estimates are significant at the 0.05 level.
*λs estimate the relative segment sizes in the finite mixture model.

4.2. Model Estimation and Comparison

We estimated the distribution of X_i using customers' responses on the degree of fit between the content of the book and their needs on a nine-point scale. Because we oversampled Group A customers, we adjusted the data as discussed in §4.1 prior to estimating the underlying distribution of X_i. A uniform distribution with mean = 5 fitted the data reasonably well, although we used the empirical distribution of X_i in the subsequent pricing application.[10]

We ran two sets of finite mixture models. In the first set, we considered consumers' purchases of print, PDF, and the bundle as distinct alternatives along with the consideration of consumers' degree of the substitutability/complementarity parameter (our proposed model). In the second set, we considered only print and PDF alternatives and the bundle purchases were considered as two separate purchases of the individual forms with no specific consideration of the substitutability/complementarity parameter (benchmark model). Because there were no specific bundle price discounts in our choice experiment, we could successfully perform such estimation. (This was one of the main reasons that we did not offer bundles at different prices.) We used the second set of results as a baseline to understand the impact of ignoring the possibility of complementarity between the two forms.

The fit of the FM models with varying number of segments ($k = 1$ through 6) considering print, PDF, and the bundle as distinct alternatives, using Bayesian information criterion (BIC) and Akaike information criterion (AIC), indicate that solutions for segments 1, 2, 3, and 6 are clearly inferior. BIC indicates that the four-segment solution is the best, whereas AIC favors the five-segment solution. We selected the four-segment solution for further consideration, as the entropy measure also favored the four-segment solution. For the second set of FM models, which consider only the print and PDF forms, a two-segment solution emerged as the best solution.[11]

4.3. Analysis and Results

Table 1 provides the estimates of the four-segment solution of the proposed model with print, PDF, and the bundle as the distinct alternatives. The results are ordered in such a way that Segment 1 has the highest preference for the printed format (also the largest segment at 40.1% share) and Segment 4 the least preference for the printed format (also the smallest segment at 10.5% share). In a relative comparison based on the

[9] Manski and Lerman's (1977) correction procedure involves weighting the different strata in the choice-based sample in the likelihood equation appropriately to ensure that the proportion of choice alternatives in the weighted sample reflect the "true" proportions of various alternative choices in the population, obtained exogenously. Such a weighted exogenous sample maximum likelihood function provides consistent estimators when using choice-based samples.

[10] There could be measurement error in X_i; therefore, we evaluated the robustness of our results to such errors. We simulated new values of X_i' with discrete uniform distribution in the range (X_i plus or minus 1) and (X_i plus or minus 2) and estimated the model over 100 simulations. With an error of plus or minus 1, the underestimation of our beta estimates was not significant. With an error of plus/minus 2, the underestimation was higher, which tended to impact the pricing policy and in a majority (77%) of the cases the policies tended to be more conservative.

[11] Details are available from the authors.

Table 2 Estimates from Finite Mixture Logit Model—No-Bundle Option

Parameters	Estimates (standard errors) Segment 1	Segment 2
β_{Print}	0.2497 (0.0518)	0.2177 (0.0397)
β_{PDF}	0.0456 (0.0136)	0.0816 (0.0107)
β_{Price}	−0.0879 (0.0213)	−0.0953 (0.0233)
λ	0.4483 (0.1312)	0 (Fixed)
Relative segment size (%)	60.0	40.0

	Log likelihood (LL)	AIC ($LL - p$)	BIC ($LL - p * \ln(N)/2$)
Model fit	−2,832.37	−2,839.37	−2,861.77

Notes. All estimates are significant at the 0.05 level.
*λs estimate the relative segment sizes in the finite mixture model.

estimates and their standard errors, Segment 3 views the forms as most imperfect, as it shows significantly ($p < 0.01$) the highest incremental preference for the bundle over and above the most preferred format, print, in that segment. For Segment 1, the estimate of $\Delta\beta$ implies that Segment 1 views the two forms as *almost-perfect* substitutes, as the incremental preference of the bundle over and above the most preferred format, print, is not different from zero ($p = 0.65$). Customers in other segments perceive the two forms to be somewhat imperfect substitutes. Segment 1 is significantly less price sensitive than the other three segments.

Table 2 provides the estimates for the optimal two-segment solution for the FM model that considers only the print and PDF as alternatives (no bundle option—the benchmark model). Both segments have higher preferences for print as compared to the PDF, whereas the difference in preferences for the forms is significantly much higher in the larger segment (Segment 1—size 60.0%) as compared to Segment 2. The price sensitivities are similar for both segments. We will use these results later to understand the impact of ignoring the form-preference interdependencies and thus the possible complementary effects between the forms.

4.4. Pricing the Product Line

There were two specific scenarios in which prices were determined for NAP's application. First, it was necessary to set prices for the situation when a printed book was already being sold at a specific price and the PDF version was being introduced for the first time. This was the case for the majority of the titles that were already being sold in printed form at the NAP website. The second situation was when the constraint of the current printed price was removed—that is, the optimal prices for both the print and PDF versions were determined simultaneously. More important, although the results showed existence of a heterogeneous customer base with widely varying preferences for formats and price sensitivities, NAP did not want to discriminate among consumers in its pricing. Thus, although customer heterogeneity was considered in developing optimal prices, they were implemented uniformly across all customers.

Given the value of the marginal costs c_1 and c_2 of the printed books and the PDF format, respectively, we solved for the optimal prices using the four-segment model. The input to the optimization consisted of the distribution of X_i, estimates of c_1 and c_2, and the other estimates from the empirical models. For the PDF version, the marginal costs were assumed to be zero. Table 3 shows the optimal prices for the product line for four titles with different marginal costs of the printed form. Three different cases are compared in this table. In column (3) the optimal prices for the two forms and the bundle are derived based on a profit maximization objective with no constraints. In column (4) the optimal prices for the PDF form and the bundle are derived, while keeping the printed price fixed at the existing level, with additional constraints on minimum penetration. This case was specifically examined because NAP did not originally plan on changing the prices of existing printed titles. In column (5) the optimal prices for print and PDF are derived under the conditions of no-bundle discount (that is, $p_3 = p_1 + p_2$). Comparing column (5) with column (3) shows that bundle discounts improve profits significantly.

In column (6) the optimal prices for the print and PDF forms are derived based on the two-segment results in Table 2, which do not consider form-preference interdependencies. However, the four-segment solution (the real heterogeneity characterizing the market) is used to derive the profits in column (6). The bundle option was not made explicitly available so if customers needed both forms they would have to purchase both formats at full price (that is, $p_3 = p_1 + p_2$). The results in Table 3 show that ignoring the aspect of imperfect substitutability or complementarity among the product forms leads to optimal prices (column (6)) that are much lower, while taking that aspect into account leads to higher optimal prices and, therefore, much higher profits (columns (3) and (4)). Interestingly, the optimal print prices in columns (5) and (6) that disregard the form-preference interdependencies are much closer to the existing printed book prices that NAP was charging before and during the choice experiment. If NAP were to leave the printed prices unchanged but have a new pricing policy only for the PDF forms

Table 3 Optimal Prices and Profits Based on the Proposed Model

(1) Marginal cost of title c_1 ($)	(2) Index	(3) Optimal prices without any constraint ($)	(4) Optimal prices with print price fixed at current levels	(5) Optimal prices with no-bundle discount ($p_3 = p_1 + p_2$) ($)	(6) Optimal prices ignoring form-preference interdependencies (no-bundle option—Table 2) ($p_3 = p_1 + p_2$) ($)
6.36	p_1^*	29.55	$23 (fixed)	24.97	25.14
	p_2^*	20.63	$20.30	16.05	18.05
	p_3^*	29.65	$29.03	41.02	43.19
	Π	8.49	$8.23	7.85	7.78
9.79	p_1^*	32.08	$25 (fixed)	27.61	27.81
	p_2^*	19.63	19.40	16.04	17.41
	p_3^*	32.15	31.55	43.65	45.22
	Π	7.64	7.37	7.10	7.03
12.73	p_1^*	34.44	$29 (fixed)	29.96	30.21
	p_2^*	19.05	$18.90	16.02	16.96
	p_3^*	34.55	$34.11	45.98	47.17
	Π	7.00	$6.88	6.55	6.54
19.58	p_1^*	39.97	$39 (fixed)	35.66	36.10
	p_2^*	17.91	$17.95	16.00	16.18
	p_3^*	40.25	$40.08	51.66	52.28
	Π	5.85	$5.84	5.53	5.49

and the bundle, then profits would be lower than if they followed the optimal pricing of column (3).

5. Policy Implementation and Aggregate Impact

5.1. Pricing Policy

From NAP's viewpoint, an optimal pricing strategy was defined as the one that ensured NAP's financial viability while increasing the dissemination of the titles beyond what was possible with only the printed form—in alignment with its dual mission. The optimal pricing policy determined using the model was in line with the objective: maximizing profits with a specified minimum penetration level. Given that customers frequenting the NAP website were already familiar with existing printed book prices, NAP decided that the price of the titles in printed form after the PDF introduction would remain the same as before the introduction for all 2,500 or so existing titles. Another reason not to raise the printed book prices and follow the higher profit pricing of column (3) in Table 3 was because the additional profits would have come at the expense of lower penetration of the product forms—a situation that NAP wanted to avoid. The optimal prices in column (4) of Table 3 (with the fixed print prices) lead to market penetration levels (not reported) that are somewhat higher than the others. It was noted at the time that if the minimum penetration-level constraint was binding in the optimization, NAP could raise prices slightly, not lose too much penetration, and realize some increased profits—higher than in column (4) but lower than in column (3).

To make the implementation as simple as possible, NAP decided to adopt a uniform pricing policy that sets prices for all titles in PDF and bundle form as a percentage of existing printed book prices. Using the model estimates, distribution of X_i, and the previous two years' distribution of prices of titles sold in printed form and their respective marginal costs in both forms as input to the optimization model, profit scenarios were created for different price levels of PDF forms and the bundle as a percentage of printed prices. Figure 2 provides an analysis of the predicted net sales from the introduction of PDF and bundle forms,

Figure 2 Determining Optimal Relative PDF Price

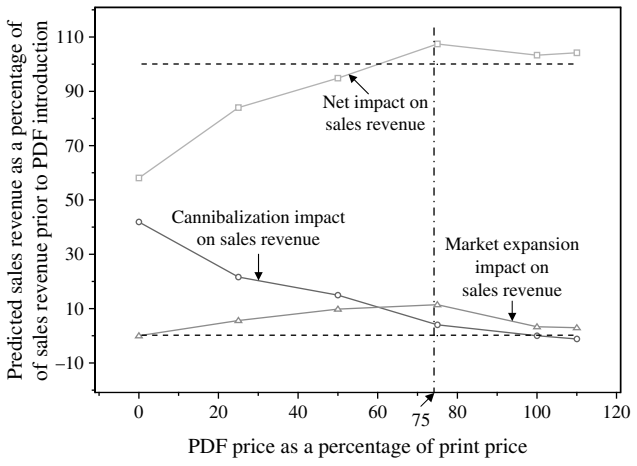


Table 4 Comparison of Daily Means Across Before-PDF and After-PDF Periods April 2002–April 2004

Variable (units per day)	Overall mean (S.D.) 4-1-02 to 4-1-04	Before-PDF mean 4-1-02 to 4-1-03	After-PDF mean 4-2-03 to 4-1-04	Percentage change between periods
All book sales by units	198.81 (132.96)	191.80	204.65 (100%)	+6.7
All book revenue	$5,306.21 ($3,309.28)	$4,936.33	$5,648.14 (100%)	+14.4
Print sales by units	178.76 (124.54)	191.80	176.12 (86%)	−13.6
Print sales revenue	$4,879.81 ($3,126.20)	$4,936.14	$4,838.14 (85.7%)	−2.0
PDF sales by units	13.23 (7.42)		13.27 (7%)	
PDF sales revenue	$299.43 ($178.56)		$300.15 (5.3%)	
Bundle sales by units	13.86 (8.15)		13.86 (7%)	
Bundle revenue	$509.99 ($303.23)		$509.99 (9%)	
Avg. no. of new titles	0.64 (1.071)	0.66 (Total 242)	0.62 (Total 226)	−6.0
No. of exhibits	1.80 (7.171)	1.47	2.12	+44.2
No. of print ads	0.35 (1.529)	0.42	0.27	−35.7
No. of mailing campaigns	0.64 (6.319)	0.71	0.56	−21.1
Off-book page	0.38 (1.795)	0.13	0.63	+384
Publicity radio interviews	0.53 (1.160)	0.52	0.54	+3.84

with prices of the PDF form expressed as a percentage of the printed price. The bundle price was kept constant at 120% of the printed price for this expositional graph. The profit-maximizing and sales-maximizing prices were the same in this case. The graph shows two impacts on the overall sales: (1) the cannibalization impact of PDF sales on existing print sales (set as the baseline of 100% in Figure 2) at PDF prices lower than the printed price, and (2) the market expansion effect from incremental PDF sales at different price levels. The profits and net sales are maximized for the PDF form at 75% of the printed prices and bundle prices at 120% of the printed prices.[12] The above pricing policy was adopted by NAP, and the PDF and bundle forms were sold in addition to printed books at the NAP website starting April 2, 2003.

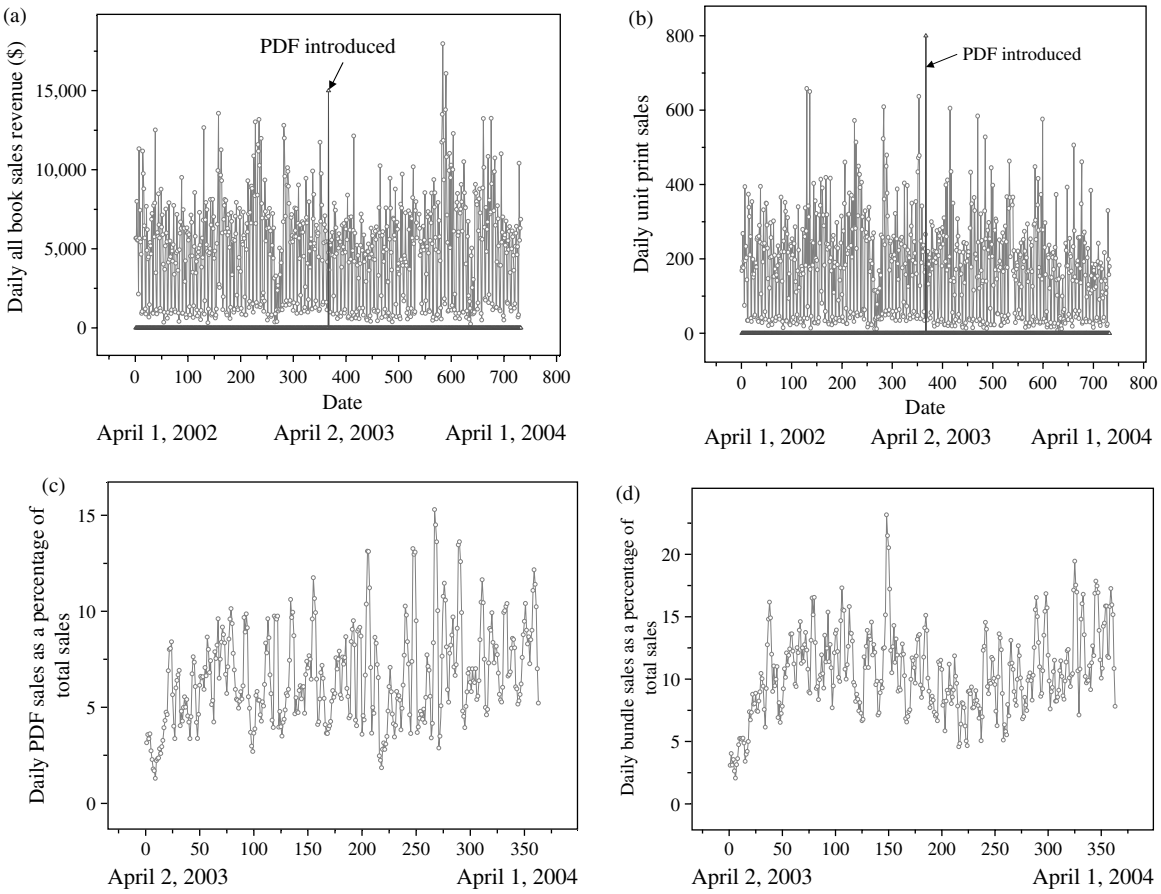
5.2. Aggregate Sales Impact: April 2003 to April 2004

One year after implementing the new pricing policy, sales data were analyzed to compare actual sales to the sales predictions that were made on the basis of the model. Table 4 provides the aggregate sales in different forms during the year (both in units and revenue), and Figure 3 (panels A through D) provides the time series of unit sales of the printed form from April 2002 to April 2004, and the PDF and bundle forms from April 2003 to April 2004. Comparing the *aggregate* predictions of print sales with actual sales, the model seemed to have performed remarkably well, especially in the case of print and the bundle. Whereas PDF sales were predicted at 10% of total sales, the actual sales came in at 7% by units sold and 5.3% by dollar sales. However, Figure 3 shows that PDF sales showed an increasing trend in the latter half of the year.

Finally, the actual increase in net sales in the Internet channel because of the introduction of the PDF format (14.4%) was almost double the predicted figures from the model (7.4%). Overall, across all channels the increase in net sales was 5%, with the offline channel (mainly institutional buyers) showing a 2% sales increase. Note that total revenue is influenced by much more than just the introduction of new forms of content. The market potential of the new titles that are introduced each year, the promotional campaigns conducted, and other such factors can drive trends.

[12] A limited sensitivity analysis of this pricing policy to variations in the estimates of the model indicated that our policy is a fairly conservative one.

Figure 3 Sales of Different Forms: April 1, 2002–April 1, 2004



In addition to promotional campaigns, a Web discount that was in effect for online printed book orders was reduced in November 2003, impacting the overall price paid. A majority of the promotional activities and patterns were not much different from the previous years, although there are significant changes in some promotional activities (see Table 4); therefore the model's prediction on market expansion cannot be accurately tested using the revenue data unless these other effects are controlled for or teased out. We discuss this further in the next section.

5.3. Other Policy Changes

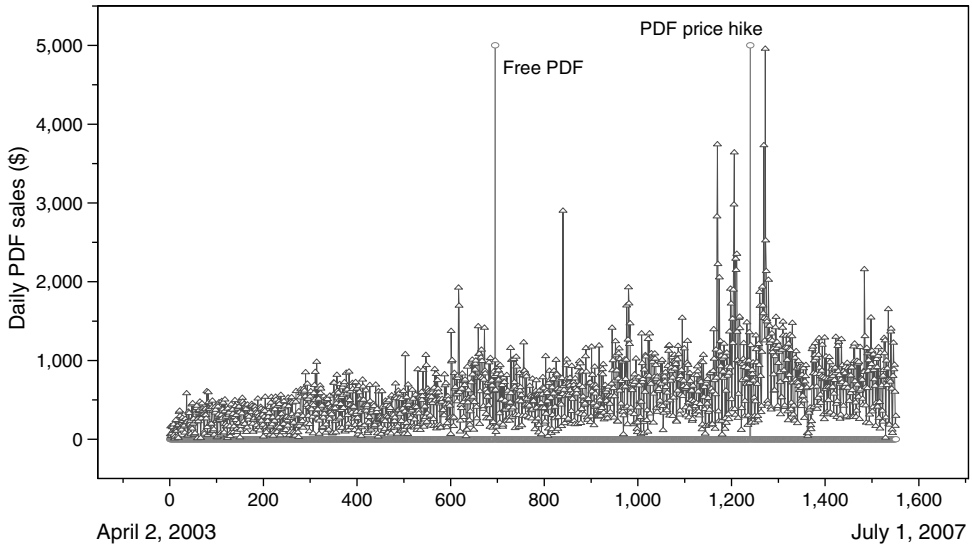
Given the positive financial experience with the PDF introduction, and the growing confidence in the model, NAP used the model to guide other policy changes, using it to predict impact on revenue and penetration and dissemination. Specifically,

- In October 2004, management decided to allow free downloads of PDF files to anyone from 144 developing countries. In February 2005, NAP decided to provide PDF files of a slow-moving category of their products called "Compass-titles" free to everyone in all countries, including the United States. A comparison of the actual sales figures before and after giving free PDFs of Compass-titles show no appreciable differences in overall sales as predicted by the model (see Table 5). (However, this difference does not control for marketing mix and pricing decisions, which is covered in §6.)
- The model suggested that NAP could charge higher prices for printed books in the presence of PDF versions (column (3) of Table 3). NAP management

Table 5 Comparison of Daily Means Across Before-Free-PDF and After-Free-PDF

Variable	Before-free-PDF mean April 2, 2003–Feb 23, 2005	After-free-PDF mean Feb 24, 2005–July 1, 2007
All book sales	194.30	186.37
All book revenue ($)	5,398.58	5,332.46
Print sales	151.72	131.03
Print sales revenue ($)	4,401.58	3,820.34
PDF sales	15.78	24.37
PDF sales revenue ($)	379.90	726.84
Bundle sales	14.94	14.86
Bundle sales revenue ($)	584.50	732.46
Number of new titles	0.57	0.61
Exhibits	2.60	2.02
Print advertisements	0.23	0.25
Mailings	0.55	0.18
Internet mailings	0.00	0.07
Off-book page	0.51	0.01
Publicity radio interviews	0.39	0.01

Figure 4 Daily PDF Sales Revenue—Time Series



began charging higher prices for printed formats for new titles beginning January 2005.

• PDF sales as a percentage of overall Web-based sales increased from 7% to 13.6% in the two periods shown in Table 5. (See also Figure 4.) This higher level of sales of PDF, indicating customers' increasing comfort with the PDF format, provided the impetus to increase the prices of PDFs relative to printed books from 75% to 85% in August 2006.

These incremental policy changes represent how the early validation of the model and the impact of initial recommendations and implementations have given NAP the necessary confidence to execute further policy changes using suggestions derived from the model.

6. Dynamic Analysis of Sales and Impact Validation

The aggregate analysis of the previous section has shown our model results to be valid; however, there are many confounds that prevent us from accurately determining the specific impact of the policies implemented on the basis of our model results. In this section we discuss our applications of persistence modeling techniques (Dekimpe and Hanssens 1999, 2004; Pauwels et al. 2004) to the daily sales time-series data to accurately estimate the impact of two specific policy changes—(1) introduction of PDF forms and the bundle, and (2) free downloads of PDF—by controlling for the impact of other variables. Introduction of new titles clearly impacts the sales of the books during the time period of analysis. Promotions such as snail mail and e-mail campaigns, publicity (news articles, book reviews, off-the-book page features, and radio interviews of authors), and exhibits at conferences also impact overall sales. In addition, the variations in the level of Web discounts and relative pricing of the product forms can have impact on overall sales. Many of these variables are also endogenous—for example, timing and depth of promotions are generally set based on the introduction of new titles, and the timing of new title releases might be related to drop in overall sales prior to new releases. Web discounts and other incentives might be directly related to an observed drop in sales; the sales levels in a particular time period are explained by past trends and the past levels of the other endogenous variables. Persistence modeling techniques allow us to take into account this endogeneity and control for the impact of these variables in determining the impact of policy changes on overall sales revenue.

6.1. Impact of the Introduction of PDF Forms

In this section, we focus on the two years' time-series data covering April 1, 2002 to April 1, 2004 that corresponds to the time frame analyzed in §5. PDF forms and the bundle were introduced on April 2, 2003; thus the duration of this analysis spans one year prior to PDF introduction and one year after PDF introduction. We use the vector autoregressive (VARX) model that captures the short-run and long-run impact of marketing variables and also allows for endogeneity among the variables.[13] Our interest focuses on two variables: overall sales revenue and printed book unit sales. The introduction of the PDF form acts as a structural break to the time-series system, and thus the impact of this structural break on the data series must be evaluated for both variables.

[13] The details of applying persistence modeling techniques are discussed elsewhere in detail (Dekimpe and Hanssens 2004, Pauwels et al. 2004); thus we will not repeat them here.

Table 6 Unit Root Test Results Summary

Variable	ADF with no trend	ADF with trend	KPSS with no trend	KPSS with trend
All book revenue	**−0.8822**	−5.1165	**1.3959**	**0.1202**
Print sales	**−0.9639**	−5.2626	**0.8286**	**0.1467**
Number of new titles	−4.13	−23.84	0.2484	0.0473
Exhibits	−16.58	−26.53	0.5098	0.1029
Print advertisements	−23.18	−24.23	0.1282	0.0379
Mailings	−27.02	−27.26	0.0837	0.0846
Off-book page	−24.94	−26.54	**1.7823**	0.0861
Publicity radio interviews	−5.48	−14.68	0.1336	0.0656

Note. Bold numbers indicate the tests show the series to be evolving.

We first examine whether the two dependent variables are mean-reverting (stationarity, which implies that the changes in them are temporary) or evolving (which implies that they have changed permanently) in the data sample under consideration. In addition, we examine the stationarity of other marketing variables used in the VARX model. We use two alternative unit root tests for this examination—an augmented Dickey-Fuller (ADF) test (without and with a deterministic trend) (Dickey and Fuller 1979) that tests whether the null hypothesis of evolution is rejected, and a KPSS test (without and with deterministic trend) (Kwiatkowski et al. 1992) that tests whether the null hypothesis of stationarity is rejected. Convergence in these tests provides additional confidence to the inference being made based on these tests. The results of these tests in Table 6 show that the overall sales revenue and printed book unit sales are both evolving series (even after allowing for the structural break because of PDF introduction), while the marketing variables satisfy the stationarity conditions. The combination of the two sales variables do not show any evidence of cointegration, even after allowing for the structural break, so there is no long-run equilibrium in these variables (see Dekimpe and Hanssens 2004).

Given the above results, the VARX model is specified as differences (Δ) for the two endogenous variables—daily overall book sales revenue and printed book unit sales—and levels of the marketing variables on each day—number of new titles, number of exhibits, print ads, mailings, and publicity. Based on the Granger causality tests[14] (Enders 2003) all of the marketing promotion variables and the Web discount are entered as endogenous variables in the model. The impact of the introduction of the PDF form and the bundle at the model-recommended relative price level is modeled as a pulse variable that takes on a value of one on April 2, 2003 (the day the PDF form was launched on the Web) and zero for other time periods. Finally, control variables included the intercept, dummy variables for the day of the week (six dummies), indicator variables to capture the launch of blockbuster titles (to account for the spikes in sales volume on specific days; see Figures 3 and 4), and a time variable to capture the growth in sales because of customers' access to broadband, comfort with technology, migration from the offline to the online channel, and experience with digital forms, all of which tend to increase with time. The specification of the VARX model is given below:

$$\begin{bmatrix} \Delta\text{ABRevenue}_t \\ \Delta\text{PrSales}_t \\ \text{NewTitles}_t \\ \text{Promotion}_t \\ \text{WebDiscount}_t \end{bmatrix} = C + \sum_{l=1}^{L} \Pi_l \begin{bmatrix} \Delta\text{ABRevenue}_{t-l} \\ \Delta\text{PrSales}_{t-l} \\ \text{NewTitles}_{t-l} \\ \text{Promotion}_{t-l} \\ \text{WebDiscount}_{t-l} \end{bmatrix} + \sum_{m=1}^{M} \text{A}_m \begin{bmatrix} \text{AddPDF}_{t-m} \\ \text{AddPDF}_{t-m} \\ \text{AddPDF}_{t-m} \\ \text{AddPDF}_{t-m} \\ \text{AddPDF}_{t-m} \end{bmatrix} + \begin{bmatrix} u_{\text{ABRevenue},t} \\ u_{\text{PrSales},t} \\ u_{\text{NewTitles},t} \\ u_{\text{Promotion},t} \\ u_{\text{WebDiscount},t} \end{bmatrix}, \tag{11}$$

where all the marketing promotion variables are represented by just one variable in the above equation for expositional clarity, L is the order of the model (determined by the Bayesian information criteria), and M the maximum lag length for the effect of the impact of PDF introduction on the endogenous variables and the error terms

$$[u_{\text{ABRevenue},t}, \ldots, u_{\text{WebDiscount},t}]' \sim N(0, \Sigma_u).$$

The estimated VARX model has significant F-statistics for overall model fit (35.25), with the adjusted R-square value of 0.677 indicating a reasonable explanatory power given that many of the variations in the individual prices of the titles and subject categories of the content is not modeled. The model also performs well on most diagnostic tests. As is the norm in dynamic modeling, we do not interpret the estimates from the VARX model directly but rather use the estimated coefficients to estimate the impulse response functions. We estimate the impulse response functions based on two forecasts—one with the

[14] All the detailed results of the tests discussed in this section are available from the authors. For the sake of brevity and clarity, only the important results are highlighted.

Table 7 Impact of Marketing Actions on Daily All Book Sales

Variable	Short-run impact	Long-run impact
Number of new titles	132.02	27.44
Exhibits	22.78	5.58
Print advertisements	—	12.18
Mailings	13.18	—
Off-book page	—	—
Publicity radio interviews	89.01	4.03
Adding PDF form	−135.67	557.34

Note. Only significant impacts at 0.05 level shown.

information set without the marketing or policy variables, and the other with the extended information set that includes the marketing variables. The difference provides us with the incremental impact of the marketing and policy variables. In our case, we have estimated the impulse response function for the units of the two variables—all book revenue and printed book unit sales—using the estimates from the difference model. The impulse response functions indicate that the dynamic effects last between 7 and 10 days for the two variables of focus and then stabilize. The details of the short-run (immediate) and long-run impacts on the overall book sales revenue are given in Table 7.

The release of a new title has a short-run (immediate) impact of \$132 on overall daily book sales while the long-run impact on daily sales is \$27 (see Table 7). Similarly, exhibits and radio publicity have a higher short-run impact on overall daily sales revenue as compared to long-run impacts. Whereas mailings have an immediate impact, print ads tend to have a long-run impact. Finally, the introduction of PDF forms and bundles at the model-recommended relative price levels has a short-run impact of a decrease of \$136 in overall daily revenues. However, the long-run impact on the overall online daily sales is an increase of \$557. This is approximately 10% of the average online daily overall sales revenue in the period from April 2, 2003 to April 1, 2004. This result is clearly more than the 7.4% increase predicted by the pricing model, and it provides good support and validation for the predictions of our pricing model.

6.2. Impact of Free PDFs of Compass Books

Persistence modeling techniques were likewise applied to time-series data of overall book sales, printed book unit sales, PDF sales, and bundle sales covering the period from April 2, 2003 to July 1, 2007, with the free Compass book PDF giveaway starting February 24, 2005. Table 5 provides the comparison of means between the two periods on the variables of focus. While the mean overall online book sales revenue remains about the same, printed book sales are clearly decreasing, and PDF sales are clearly increasing. Unit root tests reveal that the printed book and PDF sales are definitely evolving, while for the overall book sales revenue the two alternative tests—ADF and KPSS—provide different conclusions. Based on the KPSS tests, we classify overall book sales revenue as evolving and model the free PDF giveaway as the pulse shock to the series (see footnote 14).

Table 8 Impact of Free PDF Move on Daily Sales and Revenue

Variable	Short-run impact	Long-run impact
All book sales	−361.53	−18.24*
Print sales	−7.42	−2.811
PDF sales	—	6.079

*Significant at 0.10 level, others at 0.05 level, insignificant impacts constrained to 0.

Table 8 provides the impact of the implementation of free PDF downloads of Compass books on all book sales revenue, printed book unit sales, and PDF sales. The immediate impact of the free PDFs on overall daily sales revenue is negative \$362, while the long-run impact is minimal (a decrease of \$18 on daily sales revenue). This is consistent with the pricing model prediction of low impact on the overall sales revenue when the slow-moving Compass books are given away for free. This result also provides clear validation for our model. In addition, we observe that free PDFs have a long-run negative impact on printed book unit sales, while the long-run impact on PDF sales is positive (increase of 6.1 per day). This result highlights the possible complementary impact of free PDFs on different paid PDF titles.

7. Conclusions

This project clearly demonstrates the value of marketing science to practice. This engagement has allowed us to model a product line context where the various forms could range from being perceived as substitutes to being perceived as imperfect substitutes or complements, which is different from the extant literature on pricing product lines. We have used an innovative measurement methodology to determine customer preferences for the various forms, that takes advantage of the specific design of the NAP website (in determining customer interest in specific content) and the provision of the free browsing section (in measuring the fit of the content to customers' needs). We have used state-of-the-art estimation techniques not only in estimating customers' preferences but also in validating the impact of the implemented policies by controlling for the marketing mix and introduction of new titles, and teasing out the separate impact of the executed model recommendations. We have described how NAP has used the model recommendations and how we validated the recommendations not only with aggregate sales figures, but also by using dynamic modeling techniques. As a result, NAP management has had increased confidence in the application and model predictions.

From an impact viewpoint, we have described how our model recommendations formed the main determinants for NAP decisions in three important marketing policies: (1) in setting pricing policies for the PDF launch in April 2003, (2) in designing and evaluating new policies for giving away PDFs for limited product lines or customer types (October 2004 and February 2005), and (3) in making adjustments over time to the prices of printed books (January 2005) and PDFs (August 2006). The implementation of the model recommendations occurred in stages. Only after the aggregate sales figures for the year and after the PDF launch validated the model predictions did NAP embark on other decisions, growing confident in the model's validity. Our analysis using dynamic modeling techniques in this paper show that this confidence was not misplaced. Controlling for the impact of the marketing mix and introduction of new titles, the PDF launch is shown to have had the impact predicted by our model. What we have highlighted in §5 clearly shows how NAP used the model in making policy decisions that allowed it to adhere to its dual mission of increasing dissemination while remaining self-sustaining.

Throughout the multiyear engagement, this strong research-to-practice partnership and the positive business results that it engendered was essential to keeping the National Academies' leadership and its major stakeholders satisfied. All in all, NAP demonstrated great success. For example, the number of visitors to its site has increased from 5 million in 2002 to 18 million in 2006, with 120 million views of book pages. NAP has continued to be financially viable with excess revenues invested in Web environments that enhance the value to customers and serve to drive customers to both the free and for-fee content. From this perspective, our application effectively illustrates how to use marketing models smartly to aid policy decisions in online contexts: by implementing the recommendations in stages and evaluating their impact using state-of-the-art models, thereby ensuring optimal long-run results for firms.

Finally, the model and methodology we have presented have a high degree of transferability to other online content providers. Because scholarly book publishers tend to operate in local monopolies, this research is readily applicable to other publishers. NAP has been sharing the research findings widely within the $3.3 billion scholarly book sector through presentations and symposia, which was the intent behind the Andrew W. Mellon Foundation grant to support the research. In addition, newspapers, magazines, music, and video publishers, some of whom are monopolistic based on the content they publish, can use some of our research ideas, especially in the online context where the measurement schemes we propose are easy to implement. The general approach we have suggested is likely to provide an effective beginning in the quest to understand the impact of marketing policies in such online contexts.

Acknowledgments

The authors thank the editor, Area Editors Gary Lilien and John Roberts, the anonymous reviewers, Nevena Koukova, Michael Trusov, Peggy Tseng, Michel Wedel, Chuck Weinberg, and Jie Zhang for their comments and suggestions on earlier versions of the paper, and Vishal Chaudhary for help with the data analysis. The authors also acknowledge and thank the Andrew W. Mellon Foundation for their generous grant in support of this project.

References

Adams, W., J. Yellen. 1976. Commodity bundling and the burden of monopoly. *Quart. J. Econom.* **90**(3) 475–498.

Alberini, A. 1995. Optimal designs for discrete choice contingent valuation surveys: Single-bound, double-bound, and bivariate models. *J. Environ. Econom. Management* **28**(3) 287–306.

Bucklin, R. E., C. Sismeiro. 2003. A model of Web site browsing behavior estimated on clickstream data. *J. Marketing Res.* **40**(3) 249–267.

Chung, J., V. R. Rao. 2003. A general choice model for bundles with multiple category products: Application to market segmentation and pricing of bundles. *J. Marketing Res.* **40**(2) 115–130.

Cohen, M. A., E. Jehoshua, T. H. Ho. 1997. An anatomy of a decision support system for developing and launching line extensions. *J. Marketing Res.* **34**(1) 117–129.

Dayton, C. M. 1998. *Latent Class Scaling Analysis*. Sage University Papers Series on Quantitative Applications in the Social Sciences 07-126. Sage Publications, Thousand Oaks, CA.

Dekimpe, M., D. M. Hanssens. 1999. Sustained spending and persistent response: A new look at long-term marketing profitability. *J. Marketing Res.* **36**(November) 397–412.

Dekimpe, M., D. M. Hanssens. 2004. Persistence modeling for assessing marketing strategy performance. D. Lehmann, C. Moorman, eds. *Assessing Marketing Strategy Performance*. Marketing Science Institute, Cambridge, MA, 69–93.

Dickey, D. A., W. A. Fuller. 1979. Distribution of the estimators for autoregressive time series with a unit root. *J. Amer. Statist. Assoc.* **74**(366) 427–431.

Dobson, G., S. Kalish. 1988. Positioning and pricing a product line. *Marketing Sci.* **7**(2) 107–125.

Dobson, G., S. Kalish. 1993. Heuristics for pricing and positioning a product line using conjoint and cost data. *Management Sci.* **39**(2) 160–175.

Enders, W. 2003. *Applied Econometric Time Series*, 2nd ed. John Wiley & Sons, New York.

Farquhar, P. H., V. R. Rao. 1976. A balance model for evaluating subsets of multiattributed items. *Management Sci.* **22**(5) 528–539.

Green, P. E., A. M. Krieger. 1985. Models and heuristics for product line selection. *Marketing Sci.* **4**(4) 1–19.

Jain, S. 2008. Digital piracy: A competitive analysis. *Marketing Sci.* **27**(4) 610–626.

Kamakura, W. A., G. J. Russell. 1989. A probabilistic choice model for market segmentation and elasticity structure. *J. Marketing Res.* **26**(4) 379–390.

Kohli, R., V. Mahajan. 1991. A reservation-price model for optimal pricing of multi-attribute products in conjoint analysis. *J. Marketing Res.* **28**(3) 347–354.

Kohli, R., R. Sukumar. 1990. Heuristics for product-line design using conjoint analysis. *Management Sci.* **36**(12) 1464–1478.

Kwiatkowski, D., P. C. B. Phillips, P. Schmidt, Y. Shin. 1992. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *J. Econometrics* **54**(1–3) 159–178.

Lazarfeld, P. F., N. W. Henry. 1968. *Latent Structure Analysis*. Houghton Mifflin, Boston.

Liu, Y., C. B. Weinberg. 2004. Are nonprofits unfair competitors for businesses? An analytical approach. *J. Public Policy Marketing* **23**(1) 65–79.

Manski, C., S. Lerman. 1977. The estimation of choice probabilities from choice-based samples. *Econometrica* **45**(8) 1977–1988.

McConnell, K. E. 1990. Models for referendum data: The structure of discrete choice models for contingent valuation. *J. Environ. Econom. Management* **18**(1) 19–34.

Moe, W., P. S. Fader. 2004. Dynamic conversion behavior at e-commerce sites. *Management Sci.* **50**(3) 326–335.

Moorthy, K. S., I. P. L. Png. 1992. Market segmentation, cannibalization, and the timing of product introductions. *Management Sci.* **38**(3) 345–359.

Morrison, D. G. 1979. Purchase intentions and purchase behavior. *J. Marketing* **43**(2) 65–74.

Mussa, M., S. Rosen. 1978. Monopoly and product quality. *J. Econom. Theory* **18**(2) 301–317.

Pauwels, K., S. Srinivasan, J. Silva-Risso, D. M. Hanssens. 2004. New products, sales promotions and firm value, with application to the automobile industry. *J. Marketing* **68**(4) 142–156.

Reibstein, D., H. Gatignon. 1984. Optimal product line pricing: The influence of elasticities and cross-elasticities. *J. Marketing Res.* **21**(3) 259–267.

Schmalansee, R. 1984. Gaussian demand and commodity bundling. *J. Bus.* **57**(1) S211–230.

Sinha, R. K., N. Mandel. 2008. Preventing digital music piracy: The carrot or the stick? *J. Marketing* **72**(1) 1–15.

Srinivasan, V., A. K. Basu. 1989. The metric quality of ordered categorical data. *Marketing Sci.* **8**(3) 205–230.

Venkatesh, R., R. Chatterjee. 2006. Bundling, unbundling and pricing of hybrid products: The case of magazine content. *J. Interactive Marketing* **20**(2) 21–40.

Venkatesh, R., W. Kamakura. 2003. Optimal bundling and pricing under a monopoly: Contrasting complements and substitutes from independently valued products. *J. Bus.* **76**(2) 211–231.

Venkatesh, R., V. Mahajan. 1993. A probabilistic approach to pricing a bundle of products or services. *J. Marketing Res.* **30**(4) 494–508.

Zhang, J., L. Krishnamurthi. 2004. Customizing promotions on online stores. *Marketing Sci.* **23**(4) 561–578.