Journal of Applied Economics and Policy

VOLUME TWENTY SEVEN, NUMBER 1 *SPRING 2008*

Articles

A Publication of the

KEA
Kentucky Economic Association

Published by the Department of Economics
Western Kentucky University

Journal of Applied Economics and Policy

VOLUME TWENTY SEVEN, NUMBER 1 *SPRING 2008*

A Publication of the

Kentucky Economic Association

Published by the Department of Economics
Western Kentucky University

Journal of Applied Economics and Policy

The *Journal of Applied Economics and Policy* (formerly the *Kentucky Journal of Economics and Business*) is published by the Kentucky Economic Association at Western Kentucky University. All members of the Association receive the *Journal* electronically as a privilege of membership. All views expressed are those of the contributors and not those of either the Kentucky Economic Association or Western Kentucky University. The *Journal of Applied Economics and Policy* accepts submissions from all JEL classifications; however, special attention is paid to manuscripts focusing on issues relating to Kentucky and the surrounding region.

Information and the online submission process can be found at www.wku.edu/jaep. The submission fee is $20.00 and may be paid via Pay Pal using a credit card or by check made payable to the *Journal of Applied Economics and Policy*. Correspondence regarding articles submitted for publication in the *Journal* or submission fees paid by check should be sent to:

Catherine Carey, Editor
Journal of Applied Economics and Policy
Department of Economics
Western Kentucky University
1906 College Heights BLVD
Bowling Green, KY 42101
cathy.carey@wku.edu

The **Kentucky Economic Association** facilitates greater communication among state researchers from academia, government, and business, and promotes the sharing and discussion of research by economists and public policy analysts working on issues relevant to the Commonwealth. To these purposes the association holds an annual conference and publishes the *Journal of Applied Economics and Policy*. The Kentucky Economic Association's website, www.kentuckyecon.org, is maintained by Tom Creahan, Morehead State University.

Communications related to membership, business matters, and change of address in the Kentucky Economic Association should be sent to: Jeff Florea, Madisonville Community College, jeffm.florea@kctcs.edu.

Kentucky Economic Association

Officers:

President……Talina Mathews, Kentucky Office of Energy Policy
President Elect and Program Chair……..Chris Phillips, Somerset Community College
Secretary…….……Jeff Florea, Madisonville Community College
Treasurer……Michael Jones, Governor's Office for Policy Research
Editor, *Journal of Applied Economics and Policy*, Ex Officio, Catherine Carey, Western Kentucky University

Board of Directors:

2005- 2008
Stephen Lile, Western Kentucky University
Bruce Johnson, Centre College
Chris Phillips, Somerset Community College
Chuck Martie, Governor's Office for Policy Research

2006 - 2009
Ali Ahmadi, Morehead State University
Jeff Floria, Madisonville Community College
Monica Greer, E.ON-US
Bob Houston, Eastern Kentucky University

2007 - 2010
Mike Clark, Legislative Research Commission
Michael Nicholson, Transylvania University
Jeff Johnson, Sullivan University
Daniel Vazzana, Georgetown College

Kentucky Economic Association Roll of Presidents

James W. Martin	1975-76
Ray Ware	1976-77
Stephen Lile	1977-78
Dannie Harrison	1978-79
James McCabe	1979-80
Bernard Davis	1980-81
Frank Slensick	1981-82
Lawrence K. Lynch	1982-83
Clyde Bates	1983-84
Richard Sims	1984-85
Frank Spreng	1985-86
William Baldwin	1986-87
Richard Crowe	1987-88
Richard Thalheimer	1988-89
Lou Noyd	1989-90
Gilbert Mathis	1990-91
Claude Vaughn	1991-92
L. Randolph McGee	1992-93
Paul Coomes	1993-94
James R. Ramsey	1994-95
Ginny Wilson	1995-96
Bruce Sauer	1996-97
James Payne	1997-98
Mark Berger	1998-99
Martin Milkman	1999-00
Catherine Carey	2000-01
Manoj Shanker	2001-02
David Eaton	2002-03
Alan Bartley	2003-04
John T. Jones	2004-05
Brian Strow	2005-06
Tom Creahan	2006-07
Talina Matthews	2007-08

Reviewers Contributing to This Edition:

An, Lian - University of North Florida Jacksonville
Bartley, Alan - Transylvania University
Brown, Roger McCain - University of Kentucky
Cheezum, Brian - St. Lawrence University
Coates, Dennis - University of Maryland, Baltimore County
Dennis, Steve - University of North Dakota
Dynan, Linda - Northern Kentucky University
Goff, Brian - Western Kentucky University
Greer, Monica – E.ON-US
Jones, Michael - Kentucky State Government
Kostal, Thomas - Vienna University of Economics and Business Administration
Lebedinsky, Alex - Western Kentucky University
Lile, Steve - Western Kentucky University
Lin, Tin-Chun - Indiana University - Northwest
Lumpkin, Nancy - Georgetown College
Marburger, Dan - Arkansas State University
Pulsinelli, Robert - Western Kentucky University
Skidmore, Mark - Michigan State University
Strow, Claudia - Western Kentucky University
Wassom, John - Western Kentucky University
Wisley, Thomas O. - Western Kentucky University

Special thanks to Barry Nash and Bojan Savic for help with copy editing.

Conference Announcement and Call for Papers

2008 Kentucky Economic Association
Friday, October 10, 2008 -- Holiday Inn North, Lexington, KY
Guest Lecture by John Whitehead, 2008 Distinguished Economist

PROGRAM
Those interested in presenting or discussing papers, organizing a session or panel discussion, or serving as a session chair should contact:

Christopher M. Phillips
Social and Behavioral Sciences Department
Somerset Community College
808 Monticello Street
Somerset, KY 40501
Phone: (606) 451-6839
Fax: (606) 676-9065
Email: chris.phillips@kctcs.edu

The deadline for program submissions/abstracts is **September 1**.

Completed manuscripts may be submitted to the *Journal of Applied Economics and Policy* at www.wku.edu/jaep. The *JAEP* is a publication of the Kentucky Economic Association. It is listed in *Cabell's Directory of Publishing Opportunities* and can be accessed online through *EBSCOhost*.

There will be a "Best Paper Award" as well as undergraduate paper sessions and an award for "Best Undergraduate Paper." A completed paper must be submitted no later than September 1 to be considered.

A block of rooms has been reserved at the Holiday Inn North. Ask for the KEA conference rate. Please contact the hotel directly (859-233-0512).

REGISTRATION FEES:
$75 registration fee
$15 student registration fee
To register, please return this form and payment to:

Michael Jones
Office of State Budget Director
702 Capitol Avenue
Frankfort, Kentucky 40601
Phone: (502) 564-3093
Fax: (502) 564-4694
Email: jmichael.jones@ky.gov

Checks payable to: KENTUCKY ECONOMIC ASSOCIATION

Articles

Market Efficiency at the Derby: A Real Horse Race

Steven D. Dolvin
Assistant Professor of Finance, School of Business and Economics
Butler University, E-mail: sdolvin@butler.edu

Mark K. Pyles
Assistant Professor of Finance, School of Business and Economics
College of Charleston, E-mail: pylesm@cofc.edu

Abstract:

Using race data from each Kentucky Derby from 1920 to 2005, we examine whether the horse wagering market is efficient. Most prior studies in this arena test potential betting strategies that rely on posted odds, generally finding that it is extremely difficult to devise and implement any consistently successful wager (i.e., market efficiency). We extend these studies by examining underlying determinants of posted race odds, specifically focusing on the experience of auxiliary members (e.g., jockey, breeder and trainer) associated with each entrant. We find that past Derby experience is an important determinant of posted odds and that the odds-making system appears to capture relevant experience, as using this data provides no incremental information for forming market-beating wagers. Thus, we provide additional evidence supporting efficiency in horse race betting markets.

I. Introduction

By definition, market efficiency implies that relevant information is compounded quickly into posted prices, and, therefore, market participants cannot earn a consistently positive excess return simply by forming strategies based on this known set of information. The majority of market efficiency tests focus explicitly on markets for financial securities such as stocks and bonds, generally concluding that financial markets are, at least, not inefficient. However, any market that exhibits similar characteristics should also be subject to the concept of market efficiency.[1]

For example, financial markets can be characterized by the following conditions: uncertain returns, numerous profit-seeking participants, and an extensive set of available information. Obviously, these characteristics are not unique to financial markets. In fact, Ali (1998) contends that these are all relevant descriptions of the horse wagering market. Further, Quandt (1986) suggests that the horse wagering market exhibits other similarities to the financial markets. Specifically, the profitability of investors depends on both objective (skill of firm managers or horse jockeys) and subjective (what other investors or bettors think) factors.

[1] Sauer (1998) presents a comprehensive survey on the economics of wagering markets, including horse racing.

Given the similarities between horse racing and securities markets, multiple studies have examined the efficiency of wagering on horse races by attempting to identify systematic deviations that allow for the creation of consistently successful betting strategies.[2] For example, Rosett (1965, 1971), Snyder (1978), and Ali (1979) find that participants tend to overbet long-shots and underbet favorites. Further, McGlothlin (1956) and Asch, Malkiel, and Quandt (1982) find the tendency to overbet long-shots is strongest in late races.

Asch, Malkiel, and Quandt (1984) extend the above studies by examining morning line odds, finding that profits cannot be consistently earned in win betting, but it may be possible to exploit these inconsistencies in show or place betting.[3] However, Asch, Malkiel, and Quandt (1986) reverse this latter claim and conclude it is unlikely that any potential strategy could consistently generate excess profits. They conclude that bettors, as a whole, are rational and the market is efficient. This is consistent with the previous findings of Snyder (1978). Thus, the general consensus is that the horse wagering market, similar to financial markets, is efficient.[4]

The previous literature identified above typically takes posted odds as given; however, it is possible that some information generally thought to be controlled for in the odds making system may not be fully incorporated. If this is the case, then it may be possible to earn an excess return even if existing studies that focus explicitly on final posted odds (or, equivalently, the associated probability of winning) suggest otherwise. Thus, we extend existing studies by evaluating the impact of some determinants of posted odds in an effort to see if the odds fully capture certain pre-race knowledge. Specifically, we examine the effect of prior Derby experience of three of the main (non-horse) players in a race: (1) the jockey, (2) the breeder, and (3) the trainer.

We employ a unique sample with which to examine market efficiency in the horse wagering market: the Kentucky Derby. We examine all Derby entrants from 1920-2005, specifically focusing on experience of the supporting members for each horse. As suggested above, our unique data, combined with our new approach, has several potential contributions to the literature. First, we are unaware of previous work that examines specific determinants of posted race odds. Second, most prior studies have examined data sets at lesser known tracks over multiple races. We examine the most famous horse race in the United States, avoiding many potential biases that occur due to time of race differentials, different track lengths and sizes, and extent of publicity. In this, we create a more consistent data set over time, which reduces, for example, potential biases associated with early- versus late-race betting. Third, following Asch, Malkiel, and Quandt (1984) and Ali (1998), we extend the results of existing studies that examine horse racing as a test of market efficiency. Specifically, we take a closer look at the

[2] Plackett (1975) and Henery (1981) are among the first to examine the probability of winning in a horse race. Most papers use these results as a catalyst for examining market efficiency.

[3] "Win" betting refers to a wager that attempts to identify the winner of the race. "Place" ["Show"] betting refers to selecting a horse that will either win or place (2^{nd}) [win, place (2^{nd}) or show (3^{rd})]. A horse that finishes in the top three spots is also referred to as finishing "in the money."

[4] Even studies that find systematic deviations suggest that implementing the strategy is difficult, if not impossible. Thus, even if the market is not fully efficient, it is, at a minimum, transactionally efficient.

primary variable of interest in previous studies (i.e., odds) by examining how and to what extent experience influences odds.

We use a variety of methods to measure the experience of each player, subsequently applying a two-stage approach to control for potential endogeneity embedded in the data. We find that experience is indeed an important determinant in creating post odds. In fact, it appears that experience may be, other than the horse itself, the most important determinant. Further, it does appear that posted odds fully capture experience, as no significant relations remain between experience and winning (or finishing in the money) after we control for it via our two-stage approach. Our results are consistent with previous studies that find efficiency in the market, implying that posted odds (similar to financial market prices) appear to be a robust estimation of the horse's potential to win the race.

II. Data

We examine each Kentucky Derby entrant for every race from 1915 to 2005. We obtain our data from the Kentucky Derby media guide created by Churchill Downs. The data are also available online at *www.kentuckyderby.com*. Within the media guide are the Derby charts, which contain place of finish for each entrant, along with a variety of information about the race (e.g., times, post positions, and track conditions).[5]

We also compile data for each entrant's jockey, breeder, and trainer.[6] As our primary focus is to examine the impact of previous Derby experience of these auxiliary participants on the race outcome, we create several variables to measure experience. As our primary sample begins with the 1920 Derby, we use the years 1915 to 1919 as a reference point for measuring experience. Therefore, the participant is coded as having previous experience if they had participated (in the capacity judged) in any prior Derby, beginning in 1915. We examine four measures of experience for each participant. The first, *PriorDum*, is a binary variable equal to one if the horse entrant was ridden (jockey), bred (breeder), or trained (trainer) by an individual who had at least one previous entrant in the Kentucky Derby, zero otherwise.

In order to examine the extent of experience, we also create *PriorNumb*, which is defined as the number of prior entrants ridden, bred, or trained by the jockey, breeder, or trainer, respectively. If the participant had multiple horses in the same race (in the case of the breeder or trainer), both were counted as previous experience; therefore, *PriorNumb* is not necessarily the number of prior Derbies in which the breeder or trainer

[5] The authors would like to thank Ms. Cathy C. Schenck of the Keeneland Assocation for assistance locating and compiling the data used in this study.

[6] We also examine the horse owner; however, there is considerable overlap between the breeder and the owner, particularly until recent years. Therefore, we choose to exclude this participant from our primary analysis. In unreported results, however, we examine the entire study in reference to the owner and find results qualitatively equivalent to those for the breeder.

participated.[7] Further, we examine the *success* of prior experience with *PriorWin* and *PriorMoney*. *PriorWin* is a binary variable equal to one if the horse's jockey, breeder, or trainer had previously won a Derby, zero otherwise. *PriorMoney* is a binary variable equal to one if the respective participant had previously been associated with an entrant that finished in the top three, zero otherwise.

As controls, we also extract horse-type variables. Specifically, we control for geldings and fillies. Geldings are entrants that, strictly speaking, have been castrated, while fillies are female horses that have yet to reach sexual maturity. Both may have an effect on the probability of winning, as the vast majority of Derby entrants are neither fillies nor geldings.[8]

The horse's post position may also affect its potential outcome, as much attention is paid to the draw in the week prior to the Derby. For example, there have been more winners from post positions 1(12), 4(10), and 5(12) than any other position. However, this is likely influenced by the number of horses in the race (i.e., posts 1-5 would always have entrants, whereas post 20 would not). Therefore, instead of using the number of the post position as a control, we split the post position variable into three segments; (1) *inside*, (2) *mid*, and (3) *outside*. If the number of positions is divisible by 3, each segment receives an equal number of horses. For example, if there are 9 entrants, post positions 1, 2, and 3 will be characterized as having *inside* post positions, while 4, 5 and 6 will be *mid*, and 7, 8, and 9 will be *outside*.

If the number is not equally divisible, we adjust as follows. If there are n-1 entrants, where n is an equally divisible number, we split the sample into (n/3, (n/3)-1, n/3). If there are n-2 entrants, we split the sample as ((n/3)-1, n/3, (n/3)-1). In addition, we control for the field size, defined as the number of entrants in each Derby, as well as the condition of the track.[9]

Rather than using posted odds, we transform our primary independent variable into the probability of winning, which is a more consistent variable across races, particularly when different field sizes exist. We follow Ali (1998) by defining a particular entrant's probability of winning as follows:

$$probability_i = \frac{\frac{1}{1+O_i}}{\sum_{i=1}^{n}\frac{1}{1+O_i}} \tag{1}$$

[7] To examine potential nonlinear relations, in unreported results we examine *PriorNumb* squared, but we find no significance relative to the results reported.

[8] Specifically, of the 1,305 entrants, only 90 were geldings (i.e., 6.9%), while only 21 were fillies (i.e., 1.7%). Of those, only 3 geldings and 2 fillies won the derby over the period examined.

[9] The condition of the track is constant with respect to each entrant in a particular race. However, it is widely understood that certain horses tend to run better in certain conditions (e.g., sloppy or muddy tracks). This should be factored into the odds in an efficient market. Therefore, we control for this in our regressions. In unreported results, we eliminate the track condition controls from the regressions and find the results unchanged.

where O_i is each entrants posted odds. As such, for each Derby the sum of the probabilities is 1.

III. Results

We begin by examining summary statistics that measure the level and impact of the experience of each participant of interest: jockey, breeder, and trainer. We evaluate the level of experience using the metrics defined earlier (i.e., *PriorDum*, *PriorNumb*, *PriorWin*, and *PriorMoney*). Further, we give specific attention to whether experience is associated with a higher occurrence of winning (i.e., placing 1st) or finishing in the money (i.e., placing 1st, 2nd, or 3rd). Results are presented in Table 1.

Panel A reports the relation between experience and winning, where *WinDum* is defined as a binary variable equal to 1 if the entrant wins the Derby in which it was entered, zero otherwise. Panel B examines the relation between experience and finishing in the money, where *MoneyDum* is a binary variable equal to one if the entrant finished in the money in the Derby in which it was entered, zero otherwise.

Examining the results, it appears that horses ridden by jockeys with prior experience, regardless of definition, are more likely to win, as well as place in the money. We find similar results with regard to the breeder and trainer. The only exception is that breeders that had previously bred a horse that finished in the money are not necessarily more likely to win. However, taken as a whole, it appears that previous Derby experience certainly matters, at least at a univariate level. The question thus becomes whether or not this influence is fully captured in posted odds (i.e., does the market efficiently reflect information related to experience levels?).

To address this question, we extend the analysis by examining the influence of relevant variables, including experience, on posted odds. However, as defined above, rather than evaluating odds explicitly, we convert posted odds into the underlying probability of winning. Papke and Wooldridge (1996) suggest when analyzing a dependent variable whose values are constrained between zero and one, which is the case with posted probabilities, the most appropriate statistical approach is a fractional logit model. We follow this approach and present the results of the following model in Table 2:[10]

[10] For robustness, we also examine our results using a standard OLS approach. Technically speaking, there is little difference between using a logit model and traditional OLS. The main advantage of the fractional model, which is relevant to our analysis, is the predicted values are constrained between 0 and 1, while OLS structurally does not dictate this. However, in our sample, all predicted values from the OLS approach are between zero and one. Thus, even after revising the statistical approach, our results are qualitatively the same, which adds robustness to our findings.

$$Probability = \beta_0 + \beta_1 JockeyEx + \beta_2 BreederEx + \beta_3 TrainerEx + \beta_4 Gelding + \beta_5 Filly + \beta_6 Inside + \beta_7 Outside + \beta_8 FieldSz + \beta_9 GoodDum + \beta_{10} HeavyDum + \beta_{11} MuddyDum + \beta_{12} SlowDum + \beta_{13} SloppyDum + \varepsilon \quad (2)$$

Probability is the calculated probability of winning as defined in eq. (1). *JockeyEx*, *BreederEx*, and *TrainerEx* are experience variables as represented by each participant's respective *PriorDum* (Column 1), *PriorNumb* (Column 2), *PriorWin* (Column 3), or *PriorMoney* (Column 4).[11] *Gelding*, *Filly*, *Inside*, and *Outside* are as defined previously. *FieldSz* is the total number of entrants in the respective Derby. *GoodDum, HeavyDum, MuddyDum, SlowDum,* and *SloppyDum* are binary variables equal to one if the track is in each respective condition at post time, zero otherwise.[12]

Examining the results in Table 2, we find a negative relation between the size of the field and the horse's probability of winning. This is expected in that a larger field makes, presumably, for a more competitive race, and, therefore, each horse has a lower relative probability of winning. In addition, geldings are negatively associated with the probability of winning, which is consistent with the low number of geldings entered into Derbies over the sample period. The low level of participation is likely related to the perceived ineffectiveness of horses with this specific characteristic, which would be manifest in a lower probability of winning. None of the other secondary variables of interest are highly significant, although inside post position does have a moderately significant (10 percent level througout) positive relation to the probability of winning.[13]

We next turn to our primary variables of interest, i.e., the experience measures. We find each experience measure (i.e., *PriorDum*, *PriorNumb*, *PriorWin*, and *PriorMoney*) for jockeys, breeders, and trainers has a consistently significant and positive relation to the entrant's probability of winning the Derby. This indicates that odds are contingent, at least somewhat, on the experience of the auxiliary members associated with each horse. In fact, given the significance in relation to the other control variables, it appears that prior experience may be the most important determinant (other than the

[11] An obvious concern is potential correlation between the ancillary members' measures of experience used in each model. For example, if there is a high degree of correlation, then multicollinearity could result in inefficient estimates as the significance levels would be inflated. However, an examination of the correlation matrix of each set of variables indicates low levels of correlation (never exceeding .26). Therefore, it is unlikely that multicollinearity has much of an effect. Nonetheless, for robustness we redefine the models including each of the experience variables independently and find our primary results are qualitatively unchanged.

[12] The excluded variables for post position and track condition are *Middle* and *Fastdum*, respectively.

[13] It is possible that our results may be contingent upon the time period studied. For example, in the late 1980s, large horse races, such as the Derby, began to be broadcast to the public for wagering purposes (i.e., simulcast). Therefore, more people had the opportunity to place wagers on the outcome of the race. In order to examine if this is an important determinant in our study, we create variables to control for the time periods, one from 1985 to 1994 (the period where simulcasting began to gain in popularity) and the other from 1995 to 2005 (the period where simulcasting became widespread). However, including these two variables does not change the primary results. The same is true if we include a time trend variable for the entire time period. We thank an anonymous reviewer for this suggestion.

horse itself) of odds, and in-turn, the probability of winning. This remains true for all external members (i.e., the jockey, breeder, and trainer).[14]

Given that we have established a significant relation between the experience of the jockey, breeder, and trainer and the posted odds, we now examine the deeper question of market efficiency. Most previous work has attempted to do this by examining the odds in relation to the results, but they have not examined individual determinants of these odds to see if they are completely captured by the posted odds. We therefore wish to extend previous analyses by examining the second stage equation (i.e., logit) as follows:

$$Result = \alpha + \beta_1 PredProb + \beta_2 JockeyEx + \beta_3 BreederEx + \beta_4 TrainerEx + \varepsilon \qquad (3)$$

where *Result* is either *WinDum* or *MoneyDum*.[15] *PredProb* is the predicted probability of winning as calculated using the results in Table 2 (i.e., the first stage regression) for each experience measure, which allows us to control for potential endogeneity between posted odds and our experience measures. The experience variables are as defined above.

If the experience (at least as we define it) of the jockey, breeder, and trainer is fully captured in the posted odds (i.e., the predicted probability), then we expect to see no significance for the experience variables in this second stage regression. If significance remains, it is indicative of market inefficiencies, as the market has not fully processed all publicly available information and reflected such in the price (i.e., the odds) of the asset (i.e., the horse). Results are presented in Table 3.[16]

Examining Table 3, we find little-to-no significance in any of the experience measures in relation to *WinDum*. This finding indicates that, on average, experience of the auxiliary members of the horse team has been fully captured by the posted odds, and there is no consistent strategy that can be employed to "beat the market" by examining this information. Therefore, consistent with previous studies, it appears the horse wagering market is efficient.

---

[14] Obviously, the quality of the horse would be the key factor. And, it is likely that endogeneity exists in that the best horses are able to attract the most experienced jockeys. However, we have no available method for judging the quality (or experience) of particular horses, as horses only race in a single Derby. Further, we explored using a horse's prior race record; however, without a way to standardize race results across time, tracks, and competition, the use of such records are extremely limited. Thus, the experience, in addition to reflecting increased ability, may also proxy for information related to the horse itself. Further, we have no information on the horse's bloodline, which could also serve as a proxy for quality.

[15] The finishing position of each entrant is available. However, money is typically only earned on the first three horses. Exceptions are bets such as superfectas, which require the bettor to choose the first four finishing horses in order. However, our analyses do not focus on combination bets such as these, but rather on single horse bets. Therefore, we only examine horses that win or place in the money. In unreported results, we examine horses that place (i.e., finish 2nd) or show (i.e., finish 3rd) individually and find our result are unchanged in that we find no significance in the experience measures.

[16] The Logit model, which we use for examining eq. (3), possesses the independence of irrelevant alternatives property, which fits horse racing since the relative odds depend only on the characteristics of the particular horses. Further, Bacon-Shone, Lo, and Busche (1992) find that a logit model best fits this type of data.

The same is true when examining the top three finishers in each Derby, with one notable exception. We find a remaining positive and significant relation between a jockey with a prior Kentucky Derby ride and his mounted horse finishing in the money. This perhaps indicates that the jockey (who has the most in-race control of the three auxiliary members examined) can use his experience to guide a horse through the field slightly better than those with no experience. In other words, perhaps he can maximize the finish of a non-winning caliber horse, whereas an inexperienced jockey cannot.

As a simple test of the potential impact of this finding, we consider a particular scenario. Specifically, the findings suggest, for example, that a horse who is picked *ex ante* to place fourth should have a greater chance of finishing in the money if the jockey has prior Derby experience. Thus, for each race, we rank the entrants in descending order based on the calculated probability of winning. For each entrant with the fourth highest probability of winning, we identify whether the jockey has a prior Derby mount. We then test whether those with experience are more likely to finish in the money. However, the difference is insignificant, which suggests that even though there is a small statistical significance, the economic implication is small. Thus, overall, it appears that the market is, at least, transactionally efficient in relation to money horses as well as winners.

Although our primary concern has been addressed by examining market efficiency in relation to experience, for robustness we also examine the other variables used as controls in Table 2. If markets are efficient (which is the working hypothesis) then none of the other variables should have significant relations to *Result* in the second stage. Therefore, we examine the following expanded second stage model:

$$\begin{aligned} Result = \alpha + \beta_1 PredProbPriorDum + \beta_2 GoodDum + \beta_3 HeavyDum + \beta_4 MuddyDum + \\ \beta_5 SlowDum + \beta_6 SloppyDum + \beta_7 Gelding + \beta_8 Filly + \beta_9 Inside + \beta_{10} Outside + \\ \beta_{11} JockeyEx + \beta_{12} BreederEx + \beta_{13} TrainerEx + \varepsilon \end{aligned} \quad (4)$$

For parsimony, we choose to report only results from the *PriorDum* analysis. The results are presented in Table 4. Naturally, in unreported results we examine the other three experience measures as well, but the results are qualitatively identical to those reported.

As there are three additional categories of explanatory variables here, we first examine each of them separately. In column 1 we examine track condition variables, while in columns 2 and 3 we examine horse type and post position variables, respectively. In column 4, we examine all variables combined, along with the experience variables for each auxiliary player.

The results support our previous findings of market efficiency. There is a unanimous lack of significance in all explanatory variables except the predicted probability of winning as calculated from column 1 of Table 2. Again, the only exception is the prior experience of the jockey in regards to the horse finishing in the top 3. Thus, our results as a whole appear to be robust, which further strengthens the findings of previous work that concludes the horse race wagering market is efficient.

IV. Conclusion

We examine horse racing odds for the Kentucky Derby in an effort to determine whether betting markets are efficient with regard to available pre-race information. We extend previous studies by examining the determinants of posted odds, rather than taking them as given. Specifically, we examine the impact of track conditions, horse type, post position, and auxiliary players' experience on the probability of winning the Derby. We find these experience measures may be the most predictive in creating post odds for each entrant. Using multiple experience measures for the jockey, breeder, and trainer, we find a positive relationship between prior experience and the probability of winning.

We then examine market efficiency by implementing a two-stage approach, finding no significant impact on the race outcome of experience (or any other explanatory variable) remains after controlling for its impact in determining posted odds. We interpret this as further evidence, following Snyder (1978) and Asch, Malkiel, and Quandt (1986), of market efficiency in that no consistent excess return can be generated based upon publicly available information.

These results have interesting, but disappointing, implications for bettors. On a broad scale, our results are consistent with the semi-strong form of market efficiency. In other words, any information that is publicly available appears to have already been incorporated into market prices (or odds in this case), and, therefore, no excess return can be generated, no matter the effort exerted by the investor (bettor) to extract and identify such information. This does not necessarily indicate strong-form efficiency, as we have no way to define and examine the influence of private (or inside) information, which could be described in the horse racing world as a "hot tip." This would be the next logical step of examination, should one find a way to isolate and identify "private" information.

References

Ali, Mukhtar M. (1979). "Some Evidence of the Efficiency of a Speculative Market," *Econometrica* 47: 387 - 392.

Ali, Mukhtar M. (1998). "Probability Models on Horse-Race Outcomes," *Journal of Applied Statistics*, 25 (2): 221 - 229.

Asch, Peter, Burton G. Malkiel, and Richard E. Quandt (1982). "Racetrack Betting and Informed Behavior," *Journal of Financial Economics*, 10: 187 - 194.

Asch, Peter, Burton G. Malkiel, and Richard E. Quandt (1984). "Market Efficiency in Racetrack Betting," *Journal of Business*, 58 (2): 165 - 175.

Asch, Peter, Burton G. Malkiel, and Richard E. Quandt (1986). "Market Efficiency in Racetrack Betting: Further Evidence and a Correction," *Journal of Business*, 59 (1): 157 - 160.

Bacon-Shone, J., V. Lo, and K. Busche (1992). "Logistic Analyses for Complicated Bets," *Research Report 11*, Department of Statistics, University of Hong Kong.

Henery, R. J. (1981). "Permutation Probabilities as Models for Horse Races," *Journal of the Royal Statistical Society: Series B (Methodological)*, 43 (1): 86 - 91.

McGlothlin, William H. (1956). "Stability of Choices Among Uncertain Alternatives," *American Journal of Psychology*, 69: 604 - 615.

Papke, Leslie E. and Jeffrey M. Wooldridge (1996). "Econometric Methods for Fractional Response Variables With an Application to 401 (K) Plan Participation Rates," *Journal of Applied Econometrics*, 11(6): 619 - 632.

Plackett, R. L. (1975). "The Analysis of Permutations," *Applied Statistician*, 24: 193 - 202.

Quandt, Richard E. (1986). "Betting and Equilibrium," *Quarterly Journal of Economics*, 101 (1): 201 - 208.

Rosset, Richard N. (1965). "Gambling and Rationality," *Journal of Political Economics*, 73 (December): 595 - 607.

Rosset, Richard N. (1971). "Weak Experimental Verification of the Expected Utility Hypothesis," *Review of Economic Studies*, 38: 481 - 492.

Sauer, R. D. (1998). "The Economics of Wagering Markets," *Journal of Economic Literature*, 36: 2021 – 2064.

Snyder, Wayne W. (1978). "Horse Racing: Testing the Efficient Markets Model," *Journal of Finance*, 33 (4): 1109 - 1118.

Table 1: Summary Statistics

The following table presents descriptive statistics for Kentucky Derby entrants from 1920 to 2005. Panel A examines *WinDum*, which is a binary variable equal to one if the entrant placed first in the respective Kentucky Derby, zero otherwise. Panel B examines *MoneyDum*, which is a binary variable equal to one if the entrant placed first, second, or third in the respective Kentucky Derby, zero otherwise. The rows in each panel examine the percentage of jockeys, breeders, and trainers, respectively, that win (i.e., *WinDum* in Panel A) or finish in the money (i.e., *MoneyDum* in Panel B). The columns are sorted by various measures of prior Derby experience, starting with the 1915 Derby. Specifically, *PriorDum* examines whether the jockey, breeder, or trainer, respectively, had any Derby experience prior to the sample entrant. *PriorNumb* examines the extent of experience, where we examine experience in at least two previous derbies versus those that had either zero or one previous Derby. *PriorWin* examines whether the jockey, breeder, or trainer in question had won a Derby prior to the sample entrant. *PriorMoney* examines whether the jockey, breeder, or trainer had placed first, second, or third in any previous Derby. For example, the first row/first column in Panel A indicates that 8 percent of jockeys with prior Derby experience win their races, whereas only 4 percent without experience win. The difference, which is tested in the third column of each section, is statistically significant, suggesting prior experience is a significant determinate of Derby performance. The remaining entries are interpreted similarly. *t*-statistics are calculated assuming unequal variances. Data are from www.kentuckyderby.com.

Panel A:

	PriorDum			*PriorNumb*			*PriorWin*			*PriorMoney*		
	Yes	No	*t*-stat	>=2	1 or 0	*t*-stat	Yes	No	*t*-stat	Yes	No	*t*-stat
JockeyEx	.08	.04	3.31	.09	.04	3.94	.10	.06	2.48	.10	.05	3.11
n	(855)	(450)		(642)	(663)		(278)	(1,027)		(447)	(858)	
BreederEx	.09	.05	2.39	.09	.06	1.73	.14	.06	2.67	.08	.06	1.21
n	(516)	(789)		(342)	(963)		(137)	(1,168)		(276)	(1,029)	
TrainerEx	.09	.05	2.59	.11	.05	3.43	.13	.05	3.41	.11	.05	3.21
n	(611)	(694)		(392)	(913)		(194)	(1,111)		(351)	(954)	

Panel B:

	PriorDum			*PriorNumb*			*PriorWin*			*PriorMoney*		
	Yes	No	*t*-stat	>=2	1 or 0	*t*-stat	Yes	No	*t*-stat	Yes	No	*t*-
JockeyEx	.24	.12	5.43	.26	.13	6.04	.27	.18	3.13	.28	.16	4.8
n	(855)	(450)		(642)	(633)		(278)	(1,027)		(447)	(858)	
BreederEx	.25	.16	3.60	.25	.18	2.46	.33	.18	3.49	.24	.19	1.8
n	(516)	(789)		(342)	(963)		(137)	(1,168)		(276)	(1,029)	
TrainerEx	.25	.15	4.46	.27	.17	3.92	.30	.18	3.41	.27	.17	3.7
n	(611)	(694)		(392)	(913)		(194)	(1,111)		(351)	(954)	

Table 2: Stage 1

The following table presents fractional logit results from the equation:

$$Probability = \alpha + \beta_1 JockeyEx + \beta_2 BreederEx + \beta_3 TrainerEx + \beta_4 Gelding + \beta_5 Filly + \beta_6 Inside + \beta_7 Outside + \beta_8 FieldSz + \beta_9 GoodDum + \beta_{10} HeavyDum + \beta_{11} MuddyDum + \beta_{12} SlowDum + + \beta_{13} SloppyDum + \varepsilon_i$$

where *Probability* is the entrant's calculated probability of winning based upon posted odds. *JockeyEx, BreederEx*, and *TrainerEx* are the primary variables of interest and correspond to the experience measure used in each regression. Specifically, Column 1 uses *PriorDum* to measure experience, while Column 2 uses *PriorNumb* and Columns 3 and 4 use *PriorWin* and *PriorMoney*, respectively. *Gelding* is a binary variable equal to one if the entrant was a gelding, zero otherwise. *Filly* is a binary variable equal to one if the entrant was a filly, zero otherwise. *Inside* is a binary variable equal to one if the entrant's post position is one of the inside third of the starting grid, zero otherwise. *Outside* is a binary variable equal to one if the entrant's post position is one of the outside third of the starting grid, zero otherwise. The excluded category is *Mid*. *FieldSz* is the number of horses in each Derby field. *GoodDum, HeavyDum, MuddyDum*, *SlowDum,* and *SloppyDum* are binary variables equal to one if the track at post is judged to be good, heavy, muddy, slow, or sloppy, respectively. The excluded category is *FastDum*. Data are from *www.kentuckyderby.com*.

	(1) *PriorDum*		(2) *PriorNumb*		(3) *PriorWin*		(4) *PriorMoney*	
	Coef.	*p*-val	Coef.	*p*-val	Coef.	*p*-val	Coef.	*p*-val
Intercept	-1.95	.01	-1.67	.01	-1.71	.01	-1.83	.01
JockeyEx	.29	.00	.05	.01	.46	.00	.44	.00
BreederEx	.14	.03	.02	.03	.20	.07	.19	.02
TrainerEx	.35	.00	.01	.12	.33	.00	.39	.00
Gelding	-.26	.01	-.24	.01	-.24	.01	-.24	.01
Filly	.08	.66	.18	.35	.10	.63	.09	.60
Inside	.13	.10	.15	.06	.14	.07	.15	.05
Outside	.09	.24	.08	.29	.08	.34	.08	.32
FieldSz	-.08	.03	-.08	.02	-.07	.03	-.08	.02
GoodDum	.04	.70	.01	.90	.01	.94	.09	.41
HeavyDum	.23	.19	.28	.09	.21	.18	.23	.14
MuddyDum	.05	.71	.02	.89	.05	.75	.07	.61
SlowDum	-.00	.99	.06	.66	.01	.97	.06	.68
SloppyDum	.02	.86	.01	.94	.01	.91	.01	.96
N	1,305		1,305		1,305		1,305	
Pseudo. R-Sq	.1256		.1377		.1459		.1650	

Table 3: Stage 2 (Experience Measures)

The following table presents logit regression results from the equation:

$$Result=\alpha+\beta_1 PredProb+\beta_2 JockeyEx+\beta_3 BreederEx+\beta_4 TrainerEx+\varepsilon_i$$

where *Result* is *WinDum* (Panel A) or *MoneyDum* (Panel B). *WinDum* is a binary variable equal to one if the entrant placed first in the Derby, zero otherwise. *MoneyDum* is a binary variable equal to one if the entrant placed first, second, or third in the Derby, zero otherwise. *PredProb* is the predicted probability of winning as calculated using the results in Table 2 for each experience measure. *JockeyEx, BreederEx*, and *TrainerEx* are the primary variables of interest and correspond to the experience measure used in each regression. Data are from www.kentuckyderby.com.

Panel A: WinDum

	(1) *PriorDum*		(2) *PriorNumb*		(3) *PriorWin*		(4) *PriorMoney*	
	Coef.	*p*-val	Coef.	*p*-val	Coef.	*p*-val	Coef.	*p*-val
Intercept	-4.21	.00	-3.74	.00	-3.79	.00	-3.82	.00
PredProb	15.02	.00	14.98	.00	13.55	.01	13.48	.01
JockeyEx	.47	.11	-.01	.71	.09	.76	.26	.35
BreederEx	.27	.27	-.01	.81	.41	.21	-.09	.74
TrainerEx	.02	.93	.03	.07	.35	.25	.31	.27
N	1,305		1,305		1,305		1,305	
% Concordant	64.2		62.1		64.1		65.8	

Panel B: MoneyDum

	(1) *PriorDum*		(2) *PriorNumb*		(3) *PriorWin*		(4) *PriorMoney*	
	Coef.	*p*-val	Coef.	*p*-val	Coef.	*p*-val	Coef.	*p*-val
Intercept	-3.06	.00	-2.55	.00	-2.65	.00	-2.69	.00
PredProb	17.87	.00	16.89	.00	17.94	.00	18.00	.00
JockeyEx	.44	.01	-.01	.63	-.11	.58	.16	.37
BreederEx	.20	.18	-.01	.57	.29	.20	-.08	.67
TrainerEx	.02	.89	.02	.25	-.01	.95	-.05	.80
N	1,305		1,305		1,305		1,305	
% Concordant	66.0		61.8		61.1		64.1	

Table 4: Stage 2 (All Measures)

The following table presents results from the equation:

$$Result=\alpha+\beta_1 PredProbPriorDum+\beta_2 GoodDum+\beta_3 HeavyDum+\beta_4 MuddyDum+\beta_5 SlowDum+$$
$$\beta_6 SloppyDum+\beta_7 Gelding+\beta_8 Filly+\beta_9 Inside+\beta_{10} Outside+\beta_{11} JockeyEx+\beta_{12} BreederEx+\beta_{13} TrainerEx+\varepsilon_i$$

where *Result* is *WinDum* (Panel A) or *MoneyDum* (Panel B). *WinDum* is a binary variable equal to one if the entrant placed first in the Derby, zero otherwise. *MoneyDum* is a binary variable equal to one if the entrant placed first, second, or third in the Derby, zero otherwise. *PredProbPriorDum* is the predicted probability of winning as calculated with the results from column 1 in each panel of Table 2. *GoodDum, HeavyDum, MuddyDum, SlowDum,* and *SloppyDum* are binary variables equal to one if the track is in each respective condition at post time, zero otherwise. The excluded category is *FastDum*. *Gelding* and *Filly* are binary variables equal to one if the entrant was a gelding or filly, respectively, zero otherwise. *Inside* is a binary variable equal to one if the entrant's post position is one of the inside third of the starting grid, zero otherwise. *Outside* is a binary variable equal to one if the entrant's post position is one of the outside third of the starting grid, zero otherwise. The excluded category is *Mid*. *JockeyEx, BreederEx*, and *TrainerEx* are the primary variables of interest and correspond to prior experience as measured by *PriorDum*. Data are from www.kentuckderby.com.

Panel A: WinDum

	(1)		*(2)*		*(3)*		*(4)*	
	Coef.	*p-val*	Coef.	*p-val*	Coef.	*p-val*	Coef.	*p-val*
Intercept	-4.06	.00	-3.99	.00	-3.98	.00	-4.13	.00
PredProbPriorDum	19.50	.00	18.90	.00	19.17	.00	14.15	.01
GoodDum	.06	.87					.09	.82
HeavyDum	.10	.92					.06	.96
MuddyDum	.06	.87					.06	.92
SlowDum	.05	.92					.09	.87
SloppyDum	-.05	.92					-.04	.94
Gelding			-.34	.58			-.41	.51
Filly			.15	.84			.12	.88
Inside					.08	.77	.11	.68
Outside					-.24	.41	-.21	.48
JockeyEx							.48	.11
BreederEx							.29	.24
TrainerEx							.03	.90
N	1,305		1,305		1,305		1,305	
% Concordant	63.4		63.6		63.7		65.1	

Panel B: MoneyDum

	(1)		*(2)*		*(3)*		*(4)*	
	Coef.	*p-val*	Coef.	*p-val*	Coef.	*p-val*	Coef.	*p-val*
Intercept	-2.95	.00	-2.95	.00	-2.86	.00	-3.03	.00
PredProbPriorDum	22.07	.00	22.30	.00	22.16	.00	18.38	.00
GoodDum	.05	.85					.06	.81
HeavyDum	.08	.90					.07	.91
MuddyDum	.04	.90					.05	.88
SlowDum	.04	.92					.09	.79
SloppyDum	-.02	.96					-.02	.95
Gelding			.06	.86			.01	.98
Filly			-.71	.27			-.70	.28
Inside					-.06	.73	-.03	.88
Outside					-.21	.25	-.18	.32
JockeyEx							.43	.02
BreederEx							.20	.19
TrainerEx							.02	.90
N	1,305		1,305		1,305		1,305	
% Concordant	64.9		65.0		64.8		66.3	

The Demand for Higher Education at Kentucky's Public Universities, 1985 – 2001

Thomas G. Watkins

Professor of Economics, Department of Economics
Eastern Kentucky University, E-mail: tom.watkins@eku.edu

Abstract:

This paper examines the demand for higher education at Kentucky's eight public universities using annual data for the 1984 – 85 through 2000 – 01 academic years. Regression results suggest that full-time enrollment rates are generally negatively related to average public tuition as a percentage of per capita personal income, positively related to real tuition at private Kentucky universities, positively related to a measure of the wage differential between college and high school graduates, and negatively related to average earnings per job in all industries. On the other hand, part-time enrollment rates are generally negatively related to average public tuition as a percentage of per capita personal income, positively related to the statewide unemployment rate, and positively related to the wage differential measure.

I. Introduction

Prior to the 2001 – 02 academic year the Kentucky Council on Postsecondary Education (CPE) and its predecessor, the Kentucky Council on Higher Education (CHE), set resident undergraduate tuition at Kentucky's public institutions as a percentage of per capita personal income. Tuition in 1999, as a percentage of per capita personal income, was set at 13.4 percent for research institutions, 9.2 percent for the six comprehensive institutions, and 5 percent for the community and technical college system. [Woodley and Pruitt, 2006] Between the 1986 – 87 and 2000 -01 academic years, average resident tuition and fees at public universities increased at annual average rate of 6.9 percent in nominal terms and by 4.2 percent in real terms. In contrast, the GDP deflator increased at an average annual rate of 2.35 percent between 1986 and 2000. While tuition grew at a faster rate than the general price level during the period, the growth rate in tuition was much closer to that of the general price level than was true after 2001.

Between the 2001 – 02 and 2006 – 07 academic years average resident undergraduate tuition and fees increased at an average annual rate of 12.8 percent in nominal terms and by 10.2 percent in real terms at Kentucky's eight public universities. The average annual increase in the GDP deflator was only 2.7 percent over the same time period. Even though some of the annual tuition increases can be explained by relatively stagnant real state appropriations for operations at the eight universities over this period, many citizens of the Commonwealth of Kentucky fear that the recent increases in resident tuition and fees will discourage people from investing in higher education. In 2006 the CPE decided to adopt a new tuition policy that permits institutions to set

mandatory fees and tuition charges up to maximum parameters, which are determined by such factors as median family income, institutional type (research, comprehensive, or community and technical college), market factors, and benchmark institutional comparisons. [Woodley and Pruitt, 2006]

On February 12, 2007 Crit Luallen, Auditor for the Commonwealth of Kentucky, issued a report that offered some evidence that rising tuition at Kentucky public institutions since the 2002 – 03 academic year had adversely affected headcount enrollments at the eight public universities and at the community and technical college system. [Luallen, 2007] Since the Kentucky Postsecondary Education Improvement Act of 1997 sets goals for increasing workforce educational attainment by 2020, the report makes several recommendations to ensure that these goals are met. The report asks the General Assembly, the Executive Branch, the Council on Postsecondary Education, and public institutions to work together to set tuition at a level ensuring accessibility to Kentucky residents. The report recommends that reductions in tuition and increases in need-based aid be considered to improve access.

The purpose of this paper is to examine the relationship between resident undergraduate tuition and undergraduate enrollment at Kentucky's eight public universities. More specifically, the demand for higher education at Kentucky's eight public universities as a group is estimated using annual data for the academic years 1984-85 through 2000 – 01 when resident undergraduate tuition was set by the CHE or CPE. Even though complete data are available through the 2005 – 06 academic year, the time period when universities had more flexibility to set resident tuition and fees is too short to reliably measure the effect of significant increases in tuition on enrollment.[1]

The rest of this paper is presented as follows. Section II provides a brief historical perspective on enrollment at Kentucky's eight public universities. Section III offers a brief review of the literature concerning the demand for higher education. Section IV describes the statistical model and the data. Section V discusses the regression results, and Section VI offers concluding remarks.

II. Historical Perspective on University Enrollments

Full-time and part-time Fall enrollments at Kentucky's public universities for the past twenty years are briefly described in this section. From this point forward, the fiscal year will be used to describe the academic year. Fall enrollment data were collected from the annual surveys of Fall Enrollment conducted by the National Center for Education Statistics as part of the Integrated Postsecondary Education Data System.[2]

[1] Between 2001 and 2006 both undergraduate resident tuition and student aid grew at average annual rates exceeding 10 percent, which made these variables collinear. Since this time period is relatively short, the separate effects of increases in tuition and in student aid were impossible to isolate statistically. Also during this period full-time enrollment grew by almost 15 percent, and a positive relationship between tuition and full-time enrollments exists for the five-year period.

[2] Annual Fall enrollment data is available from Fall 1984 (fiscal year 1985). The Council on Postsecondary Education has headcount and full-time equivalent enrollment data for Fall 1982 and Fall 1983, but cannot provide full-time and part-time undergraduate enrollments for both genders or by gender for these years.

Figure 1 illustrates full-time enrollment and the Commonwealth's population aged 18 to 24 years between 1985 and 2006 and part-time enrollment between 1987 and 2006.[3] The 18 – 24 population, the age group composed of "traditional" college students, decreased between 1985 and 1991, was relatively stable until 2000, increased slightly between 2001 and 2003, and then decreased after 2003. Overall, this population group decreased by 15.9 percent or from a high of 456,670 persons in 1985 to a low of 384,158 persons in 2006. In contrast, full-time enrollment at the public universities increased between 1985 and 1993, was relatively stable between 1994 and 2001, and increased after 2001. Full-time enrollment increased from 55,026 in 1985 to 66,657 in 2001 and then to 76,355 in 2006. Part-time enrollment, on the other hand, increased from 1987 to 1992, decreased between 1993 and 1999, was relatively stable between 2000 and 2003, and then decreased after 2004. Overall, part-time enrollment was stagnant over the period.

Figures 2 and 3 illustrate full-time and part-time enrollments and the population aged 18 to 24 years for males and females, respectively, between 1985 and 2006.[4] Overall, in Figure 2 the male population between 18 and 24 years decreased from a high of 231,217 men in 1985 to 196,967 men in 2006. Male full-time enrollment increased from 27,590 to 33,769 mean over the period, while male part-time enrollment decreased slightly from 7,957 to 6,959 men. As Figure 3 illustrates, the female population between 18 and 24 years decreased from a high of 225,453 women in 1985 to 187,191 women in 2006. Female full-time enrollment increased from 27,436 to 42,586 women over the period and increased by more than male full-time enrollment over the period. Female part-time enrollment decreased slightly from 10,928 in 1987 to 10,613 in 2006. Part-time enrollment for both males and females decreased slightly over the period.

Figures 4 and 5 illustrate full-time and part-time enrollment rates, respectively. Enrollment rates are computed by dividing enrollment for a group by the group's population. As shown in Figure 4, the total full-time enrollment rate increased between 1985 and 1993, decreased unevenly until 2001, and then increased. The male and female enrollment rates exhibit a similar pattern, but the female enrollment rate has been increasing since 1997. Interestingly, the full-time enrollment rate of men and women combined has increased by about 3.6 percentage points since 2001, even though this is the same period when nominal resident undergraduate tuition and fees were increasing at an average annual rate of 12.8 percent. Part-time enrollment rates, shown in Figure 5, generally increased between 1985 and 1992, decreased between 1993 and 1999, and have shown little change since 2001.

Overall, these descriptive statistics offer a somewhat positive picture of university enrollments. In spite of a generally decreasing trend in the population of traditional college students, full-time enrollments and full-time enrollment rates for men and women combined, for men, and for women have increased. Part-time enrollments and part-time enrollment rates, however, have shown little change over the period.

[3] Due to measurement errors in part-time Fall enrollments in 1985 and 1986, these years were excluded.
[4] Male and female enrollments are not available for the 1999 – 00 academic year, since the National Center for Education Statistics has not released institutional data for that year.

III. Literature Review

Numerous studies have examined the demand for higher education. Most recent studies have generally employed regression analysis to estimate models based upon the theory of demand [Becker, 1990], and most have assumed the supply of higher education is perfectly elastic. The common approach is to use enrollment or enrollment rates to measure quantity demanded. The models commonly attempt to explain enrollment or enrollment rates with explanatory variables measuring tuition and fees, the prices of substitutes, income, student aid, the return to higher education, and the opportunity cost of enrolling in higher education.

Among the papers most relevant to this study, Hopkins [1974] used a cross-section in 1963 – 64 to explain public enrollment rates. The results suggested that public enrollment rates are negatively related to net public tuition, negatively related to the proximity of private institutions, negatively related to high income incidence, and positively related to the educational attainment of the head of a family. Lehr and Newton [1978] studied freshman enrollments in Oregon and found that freshman enrollment is negatively related to real tuition and positively related to per capita income, annual unemployment, the number of people aged 18 – 21 in the Armed Forces, and the number of high school graduates. Leslie and Brinkman [1987] conducted a meta-analysis of 25 studies and concluded that enrollment is consistently negatively related to tuition. McPherson and Schapiro [1991] examined enrollment rates for different income groups at both public and private institutions and found that changes in the net cost of higher education negatively affected enrollments of low-income whites, but did not adversely affect enrollments of more affluent students. Wetzel, O'Toole, and Peterson [1998] studied the sensitivity of enrollment yields of white and black students to changes in the net cost at a single public university. Heller [2001] examined public college enrollment rates and generally found that enrollment rates are negatively related to public tuition and positively related to state aid to students. Berger and Kostal [2002] estimated a simultaneous equations model to estimate the demand for and supply of public higher education using state level data on enrollment rates. On the demand side, their results suggest that public enrollment rates are negatively related to public tuition and positively related to average wages of production workers and the educational attainment of the population. Overall, most studies have shown that enrollment is negatively related to tuition, and in most enrollment demand is inelastic.

IV. Model and Data

If the market for higher education is competitive, student tuition and enrollments are simultaneously determined through the interaction of the market demand for and market supply of higher education. In this case student tuition is an endogenous variable, and the demand for higher education would need to be estimated as part of simultaneous equations model, like that estimated by Berger and Kostal [2002]. However, a simultaneous system was not estimated here for two reasons. First, since the CHE and CPE fixed full-time resident undergraduate tuition at Kentucky public universities as a percentage of state per capita income prior to 2001, this policy made undergraduate

tuition exogenous until 2001. Second, state appropriations to public institutions, federal appropriations, grants, and contracts to public institutions, and enrollments at the Commonwealth's private institutions were collected to explain the quantity supplied of public higher education and to estimate a simultaneous system using two-stage least squares. A Hausman specification test was then employed to compare the models estimated by ordinary least squares (OLS) and those estimated by two stage least squares. The null hypothesis that OLS estimation is the correct specification could not be rejected.

The demand for public higher education was therefore estimated directly using annual data between 1985 and 2001[5] with the following model:

$$Enrollment\ rate = \beta_1 + \beta_2 FTPCI + \beta_3 RPRVT + \beta_4 UR + \beta_5 WD + \beta_6 REPJ + \varepsilon \quad (1)$$

In this equation, the annual quantity demanded of higher education at the Commonwealth's eight four-year universities is the undergraduate enrollment rate of different groups of students. As described earlier and as shown in Table 1, the undergraduate enrollment rate is the percentage of total state population between the ages of 18 and 24 enrolled in public institutions as full-time or part-time undergraduates. Enrollment rates for male and female students who attend full-time and part-time are also measured in a similar manner.

As shown in Table 1, the explanatory variables include most of the important variables influencing an individual's decision to enroll in higher education. Real tuition (FTPCI) is average full-time resident undergraduate tuition and fees as a percentage of per capita personal income. Tuition is the "sticker" price of higher education, since tuition, net of scholarships, is not available. Since tuition was set as a percentage of per capita personal income until 2001, a linear relationship existed between tuition and per capita personal income over time. To control for this linear relationship, real tuition is expressed as a percentage of per capita personal income and can be interpreted as a measure of affordability. Theory suggests that an individual is less likely to enroll as tuition increases, so a negative coefficient is expected.

Real full-time undergraduate tuition at four-year Kentucky private institutions (RPRVT) measures the inflation-adjusted "sticker" price of enrollment substitutes within Kentucky using the personal consumption expenditures deflator to adjust for changes in the price level.[6] While some students consider both public and private out-of-state institutions as good substitutes for Kentucky public universities, there was no objective way to include these substitutes without more detailed information about student

[5] Demand can be estimated using annual data from 1985 to 2001 for full-time enrollment rates and using annual data from 1987 to 2001 for part-time enrollment rates.

[6] Undergraduate tuition at private institutions was also measured as a percentage of per capita income, like the public tuition variable, but this variable proved less effective in explaining the enrollment rates.

choices.[7] As real private tuition increases, an individual is more likely to enroll in a public university, so a positive coefficient is expected.

Since an individual may choose to work full-time rather than enroll in higher education, the unemployment rate (UR) measures the likelihood that an individual will find acceptable employment in lieu of enrollment. One normally expects an individual to be more likely to enroll as the unemployment rate rises. However, a negative relationship is possible if the unemployment rate captures the economic status of the student's parents.

The theory of human capital suggests that an individual is more likely to invest (or enroll) in higher education as the returns to higher education increase. The returns to higher education can be measured as the ratio of the wages of college graduates with a baccalaureate degree to the wages of high school graduates. Unfortunately, annual wage data for individual states is not available for the entire period. Annual data on wage disbursements and employment are available from the Bureau of Economic Analysis by SIC code through 2000 and by NAICS code after 2000. With this information earnings per job can be calculated for different industries. However, this data must be used cautiously for two reasons. First, the NAICS industry definitions cannot be perfectly matched to SIC industry definitions, and relatively few industry definitions are similar under the two classifications. By carefully comparing total employment and earnings per job for different industry definitions under the two classification schemes, some industries appear comparable enough to use. Second, earnings per job in a specific industry is admittedly an imperfect measure of the wages earned by college and high school graduates, since each industry employs people with different academic credentials. If some industries are more likely, on average, to hire college graduates, then earnings per job in those industries are more indicative of the wages of college graduates. On the other hand, if some industries are more likely, on average, to hire high school graduates, then earnings per job in those industries are more indicative of high school graduates. Given these considerations the weighted average real earnings per job in health services, at depository and non-depository institutions, and at security and commodity brokers was chosen to measure the real wage of college graduates. The weighted average real earnings per job in retail apparel, building supply, and general merchandise stores was chosen to measure the real wage of high school graduates. The wage differential (WD) was then the ratio real earnings per job for college graduates to the real earnings per job for high school graduates, as defined above. Since an individual is more likely to enroll in higher education as the wage differential increases, a positive relationship is expected.

The most significant cost of enrolling in higher education is the opportunity cost, as measured by the foregone wage. To measure the foregone wage, real earnings per job in all industries in Kentucky (REPJ) was calculated using wage disbursements and

[7] Students may also consider public and private 2-year institutions as good substitutes for public universities. However, since tuition at Kentucky's community and technical college system was also set in a manner similar to that of the universities, community college tuition proved to have a strong collinear relationship with university tuition and was not included as an explanatory variable.

employment data from the Bureau of Economic Analysis. Given the relatively low percentage of college graduates in the Commonwealth's population, this variable is likely to provide a reasonable measure of the expected earnings of high school graduates, since most jobs in the Commonwealth are filled with individuals having less than a college education. Also, this variable can be computed without worrying about inconsistent SIC and NAICS industry definitions. Wage disbursements were adjusted for inflation by using the personal consumption expenditures deflator. As the real foregone wage increases, an individual is less likely to enroll.

In addition to the variables defined in Table 1, other explanatory variables were collected to measure need-based and merit aid to students and to measure demographic characteristics of the population. Since theory suggests that student aid is likely to increase the demand for higher education, student aid data were collected for the period. Kentucky provided only need-based aid to college students prior to 2000, but after that provided both need-based and merit aid. When student aid is adjusted for inflation and divided by the population aged 18 to 24 years, real student aid per capita was relatively low between 1985 and 2001, ranging from $17 in 1985 to $163 in 2001. Real per capita student aid appeared to be collinear with one or more other independent variables, particularly the ratio of tuition to per capita personal income, and was therefore not included in the final models.

The percentages of blacks, Hispanics, and Asian-Pacific Islanders in the state population were calculated to measure demographic characteristics. Since the Commonwealth had relatively few citizens categorized as Hispanic or Asian-Pacific Islanders during the relevant period, neither percentage proved to be a meaningful explanatory variable. While the Commonwealth's black population is larger than the other two population groups, this percentage exhibited little variation over the period and was ineffective as an explanatory variable. As a result, the percentage of blacks in the population was not included.

The data used to estimate the demand equation were collected from three different sources. The statewide unemployment rate, the personal consumption expenditures deflator, and wage disbursements and employment by industry were collected from the Bureau of Economic Analysis. Fall enrollment data at four-year public institutions in Kentucky and mean nominal tuition at the Commonwealth's private universities were collected from the National Center for Education Statistics. Nominal average tuition and fees at Kentucky's public universities were collected from the Council on Postsecondary Education. The descriptive statistics for all variables are shown in Table 2.

V. Statistical Results

Equation (1) was estimated using the combined full-time enrollment rate for men and women, the full-time enrollment rate for men, the full-time enrollment rate for women, the part-time enrollment rate for men and women, the part-time enrollment rate for men, and the part-time enrollment rate for women. Ordinary least squares were used

to estimate all relationships. For all equations robust standard errors were used for hypothesis testing of parameter estimates. The results assuming a linear model are shown in Table 3.

Average real tuition as a percentage of per capita personal income (RTPCI) was consistently negative and significant at the .10 level or less for all equations.

The coefficient on real tuition at private universities (RPRVT) was positive, as expected, in five of six equations, but only significant at the .01 level for the three full-time enrollment rates. These results suggest that full-time students are more sensitive than part-time students to real tuition charged at private institutions. Several explanations for these results are plausible. If employers of part-time students reimburse them for tuition expenses, then they might be less sensitive to the real tuition charged by private universities. Alternatively, if part-time students are more likely to have the financial responsibilities of a family, then they may not consider private institutions a good substitute for lower-priced public institutions, all else constant.

The statewide unemployment rate is consistently positive as expected, but only significant at the .05 level in the equations measuring part-time enrollment rates and at the .10 level in the equation for female full-time enrollment. Part-time students and female full-time students are more likely to enroll as the unemployment rate rises.

The coefficient on the wage differential measure is consistently positive and significant across all equations. For all but male full-time students, this coefficient is significant at the .01 level. Apparently, the returns to higher education consistently motivate both full-time and part-time students to enroll.

The coefficient on the real earnings per job in all industries is negative in all equations, but only significant in the combined full-time enrollment rate of men and women equation and the male full-time enrollment rate equation. Again, these results seem plausible. Since many part-time students work full-time, employed part-time students are not incurring the same opportunity cost as full-time students. Foregone earnings for part-time students therefore are less important.

Overall, each equation explains a significant percentage of variation in enrollment rates, and the F-test for each equation suggests that for each equation the composite null hypothesis that all coefficients on the explanatory variables are zero is rejected at the .01 level. An examination of the residuals for each equation suggested no serious problems with serial correlation. The Durbin-Watson Statistic for each equation also suggests that serial correlation cannot be conclusively supported.

To measure the elasticity of each enrollment rate given a change in each explanatory variable, the value of each variable, other than the unemployment rate, was converted to its natural logarithm. Table 4 shows the results of estimating each equation as a log linear model. Overall, the results are very similar to those illustrated in Table 3.

Given the Commonwealth's concern about rising tuition at public universities, the elasticity of the enrollment rate given a change in tuition as a percentage of per capita personal income is perhaps the most interesting result in Table 4. This elasticity is consistently larger than -1 across all enrollment rates and significant for all enrollment rates. While the elasticity of demand is inelastic for both full-time and part-time students, part-time students appear to be slightly more sensitive to changes in tuition. For example, the elasticity for full-time enrollment for men and women combined suggests that a one percent increase in the ratio of public tuition to per capita personal income reduces the full-time enrollment rate by .29 percent, whereas the elasticity for part-time enrollment for men and women combined suggests that a one percent increase in the ratio of public tuition to per capita personal income reduces the part-time enrollment rate by .53 percent.

Another interesting result in Table 4 is the elasticity of enrollment demand given a change in the wage differential measure. For all full-time and part-time enrollment rates, this elasticity is positive and significant. However, the elasticities for part-time students exceed those of full-time students. Part-time students are again more sensitive than full-time students to changes in the wage differential. This result seems plausible, since many part-time students are employed and have more opportunities to observe directly the benefits of higher education in the work place.

As for the remaining elasticities in Table 4, the elasticity of enrollment demand given a change in real private tuition is positive, significant, and less than one for each of the three full-time enrollment rates. The elasticity of demand given a change in the unemployment rate is positive, significant, and less than one for each of the part-time enrollment rates and for the female full-time enrollment rate. Finally, the elasticity of enrollment demand given a change in the real earnings per job in all industries is negative, significant, and approximately one for the male full-time enrollment rate.

Each log-linear equation explains a significant percentage of variation in enrollment rates, and the F test supports this conclusion. The Durbin Watson Statistic for these equations suggests that serial correlation cannot be conclusively supported.

VI. Conclusion

The results presented in this paper suggest that both full-time and part-time students were relatively insensitive to changes in public tuition as a percentage of per capita personal income between 1985 and 2001. However, increases in public tuition do significantly reduce enrollments for most students, and the negative effect on enrollments is larger for part-time students. Since the CPE hopes to encourage many individuals with some college work to finish their undergraduate education and many may do so as part-time students, the results suggest the public universities need to carefully consider how tuition increases negatively affect part-time enrollments.

While these results are relevant to the Commonwealth's higher education officials and stakeholders, they are not complete. As the introduction suggested, the

Commonwealth's public universities have increased tuition at higher average annual rates than was true for the period covered in this study. While the time period since 2001 is insufficient to measure reliably the effects of more recent tuition increases, total full-time enrollment and the total full-time enrollment rate at the eight public universities as a group have increased significantly since 2001. To some extent, increases in full-time enrollment during a period of significant tuition increases may be explained by increases in student aid since 2000. The Commonwealth provided only need-based aid before 2000, but has since provided both need-based and merit aid. Total real aid to students grew at an average annual rate of 7.2 percent between 1985 and 2000 and at an average annual rate of 20.3 percent after 2000. Since both real tuition and total real student aid increased significantly between 2001 and 2006, the variables were collinear, making it impossible to reliably separate the effects of the two variables. In the future as more data becomes available, future research may be able to address this issue more fully.

References

Becker, W. E. (1990). "The demand for higher education." In S. A. Hoenack and E. L. Collins, *The economics of American universities* (pgs. 155 -265), Albany, NY: State University of New York Press.

Berger, M. & T. Kostal (2002). "Financial Resources, regulation, and enrollment in US public higher education," *Economics of Education Review*, 21 (2): 101 – 110.

Heller, D. E. (1999). "The effects of tuition and state financial aid on public college enrollment," *The Review of Higher Education*, 23 (1): 65 – 89.

Hopkins, T. D. (1974). "Higher education enrollment demand," *Economic Inquiry*, 12 (March): 53 – 65.

Lehr, D, K. & J. M. Newton (1978). "Time series and cross-sectional investigations of the demand for higher education," *Economic Inquiry*, 16 (July): 411 – 422.

Leslie, L. L. & P. T. Brinkman (1987). "Student price response in higher education: the student demand studies," *The Journal of Higher Education*, 58 (2): 181 – 204.

Luallen, C. (2007). "Recent Kentucky tuition increases may prevent the achievement of the commonwealth's 2020 postsecondary education goals." Auditor of Public Accounts, Commonwealth of Kentucky. Retrieved June 30, 2007 from www.auditor.ky.gov/Public/Audit_Reports/Archive/2007TuitionBriefing-Performance-PR.htm.

McPherson, M. S. & M. O. Schapiro (1991). "Does student aid affect college enrollment? New evidence on a persistent controversy," *The American Economic Review*, 81 (1): 309 – 318.

Wetzel, J., D. O'Toole, & S. Peterson (1998). "An analysis of student enrollment demand," *Economics of Education Review*, 17 (1): 47 – 54.

Woodley, S. & J. Pruitt (2006). "Tuition Policy," Council on Postsecondary Education Commonwealth of Kentucky. Retrieved June 30, 2007 from www.cpe.ky.gov.

Figure 1: Undergraduate Enrollment and Population Between 18 and 24

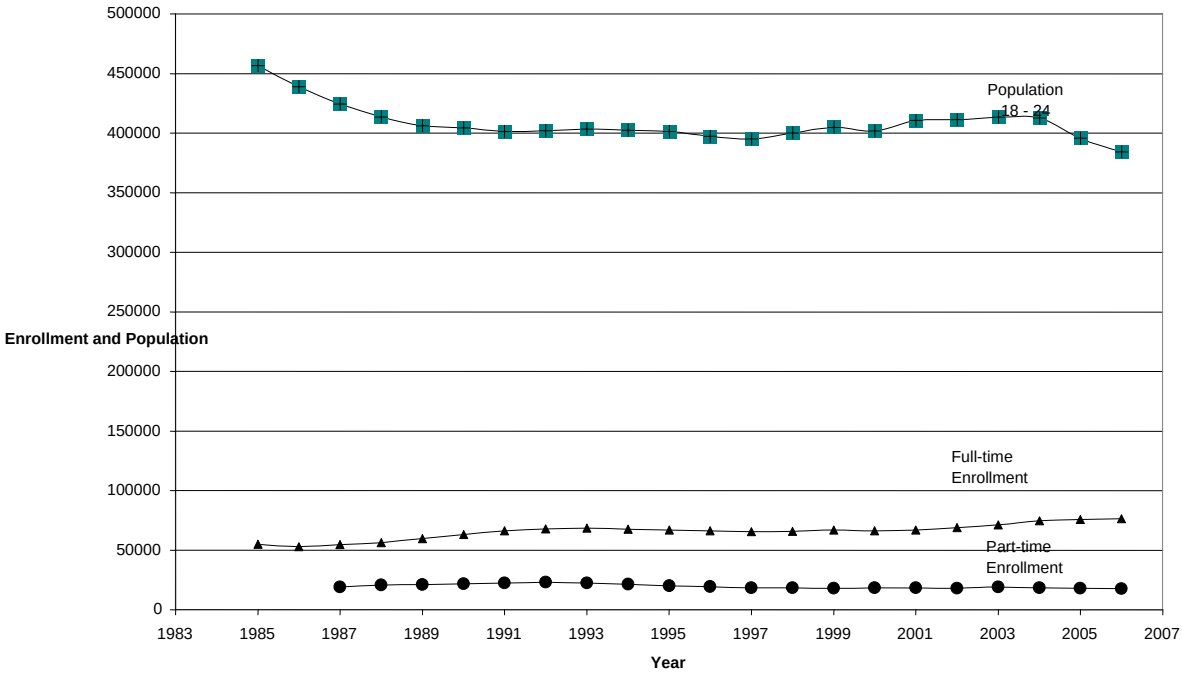


Figure 2: Male Undergraduate Enrollment and Population Between 18 and 24

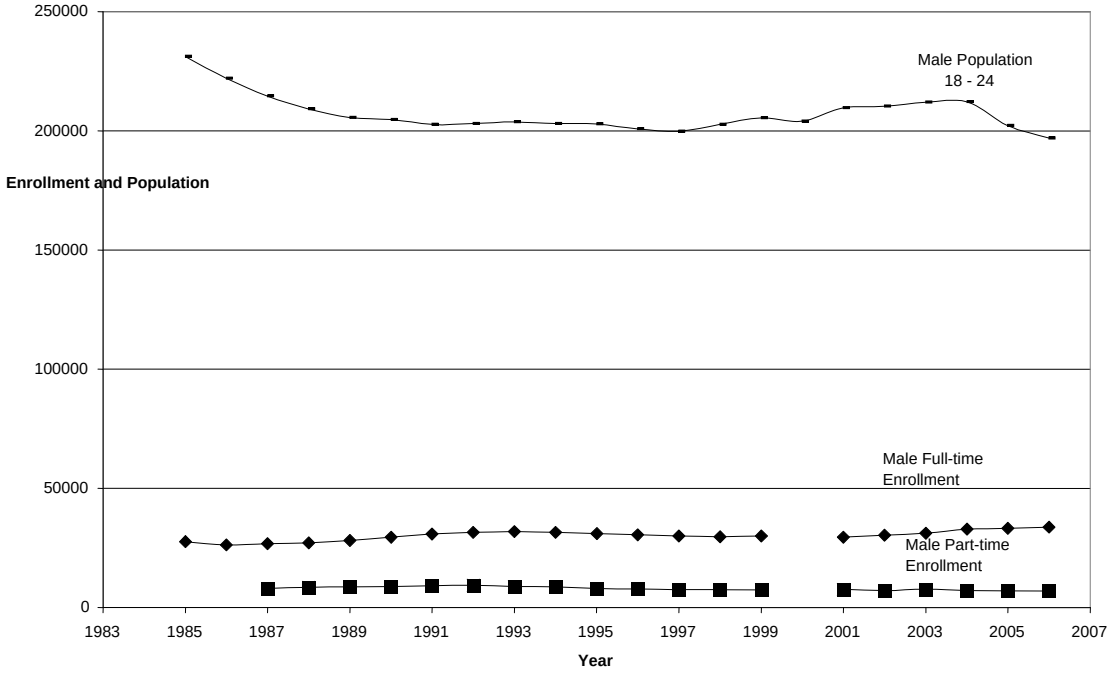


Figure 3: Female Undergraduate Enrollment and Population 18 - 24

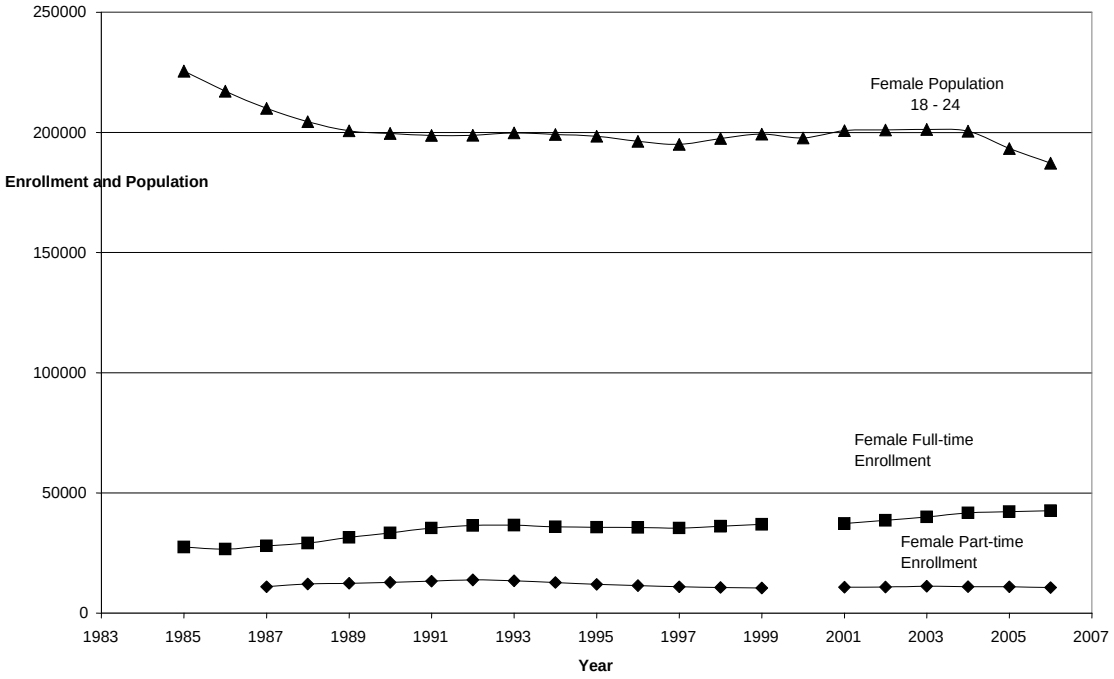


Figure 4: Full-time Enrollment Rates

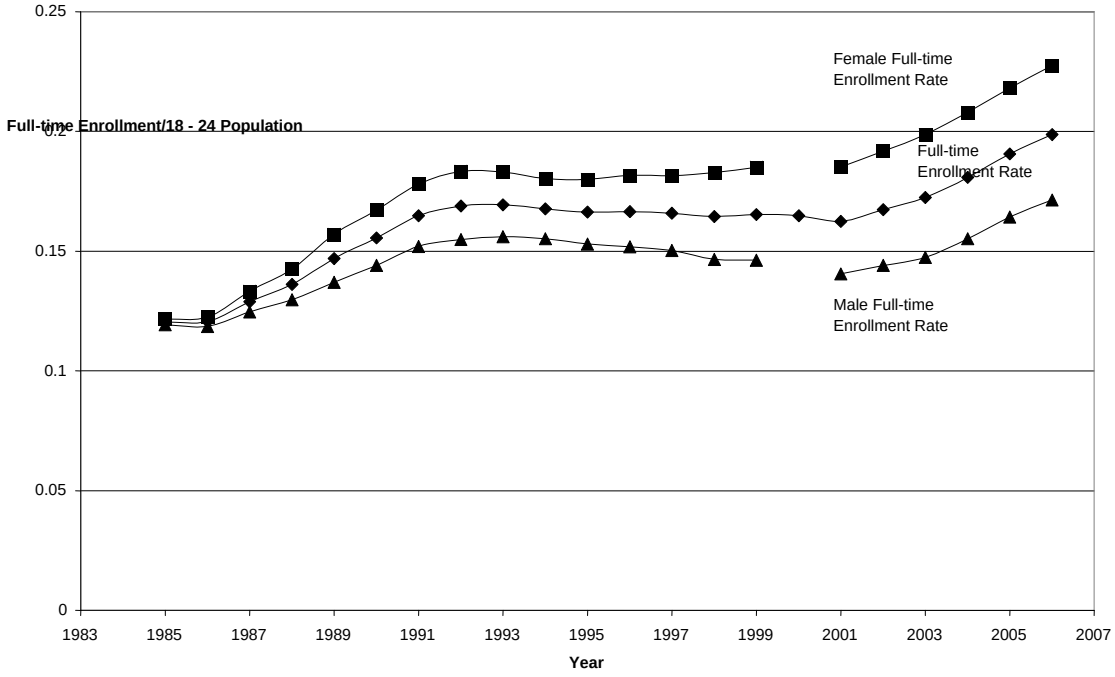


Figure 5: Part-time Enrollment Rates

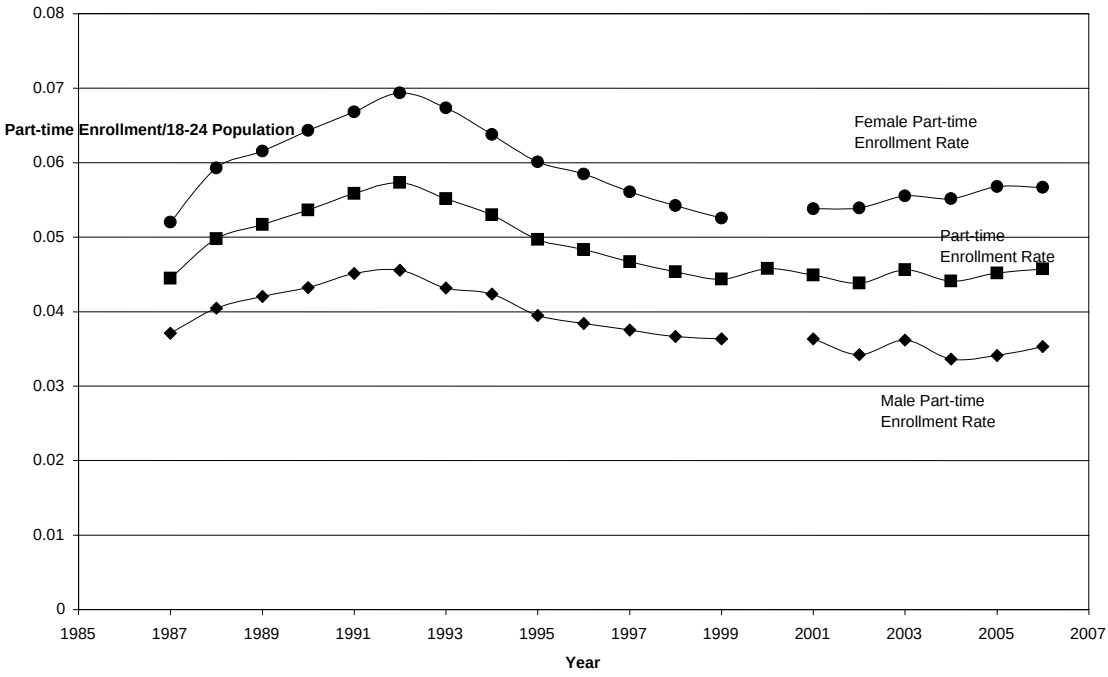


Table 1: Variable Definitions

Variable	Definition
FTRATE	Percentage of the population between 18 and 24 years of age enrolled as full-time undergraduates.
MFTRATE	Percentage of the male population between 18 and 24 years of age enrolled as full-time undergraduates.
FFTRATE	Percentage of the female population between 18 and 24 years of age enrolled as full-time undergraduates.
PTRATE	Percentage of the population between 18 and 24 years of age enrolled as part-time undergraduates.
MPTRATE	Percentage of the male population between 18 and 24 years of age enrolled as part-time undergraduates.
FPTRATE	Percentage of the female population between 18 and 24 years of age enrolled as part-time undergraduates.
FTPCI	Average full-time undergraduate tuition at Kentucky public 4-year institutions divided by Kentucky per capita income, measured in percentage points.
RPRVT	Real full-time undergraduate tuition at 4-year Kentucky private institutions, measured in $1000 of dollars.
UR	Kentucky statewide unemployment rate.
WD	Ratio of the weighted average of earnings per job in health services, financial intermediaries, and security and commodity brokers to the weighted average earnings per job in retail apparel, building materials, and general merchandise stores in Kentucky.
REPJ	Real earnings per job for all industries in Kentucky, measured in $1000 of dollars.

Table 2: Summary Statistics

Variable	Mean	Standard Deviation
FTRATE	15.5	1.74
MFTRATE	14.3	1.29
FFTRATE	16.7	2.34
PTRATE	4.97	.448
MPTRATE	4.03	.33
FPTRATE	5.99	.577
FTPCI	9.6	.888
RPRVT	7.1	1.79
UR	6.4	1.53
WD	2.00	.097
REPJ	25.84	1.46

Table 3: Regression Results for Linear Model (Standard Errors in Parentheses)

Independent Variables	Dependent Variables					
	FTRATE	MFTRATE	FTRATE	PTRATE	MPTRATE	FPTRATE
FTPCI	-.517* (.157)	-.306*** (.142)	-.692* (.188)	-.296** (.098)	-.215** (.085)	-.299** (.112)
RPRVT	1.31* (.152)	1.10* (.123)	1.52* (.20)	.043 (.09)	-.002 (.087)	.052 (.103)
UR	.339 (.213)	.282 (.21)	.44*** (.23)	.218** (.09)	.17** (.07)	.334** (.105)
WD	9.48* (2.6)	5.54*** (2.54)	13.93* (2.64)	7.17* (1.76)	5.14* (1.31)	10.16* (1.99)
REPJ	-.497** (.20)	-.735* (.19)	-.26 (.21)	-.08 (.13)	-.07 (.11)	.091 (.15)
CONSTANT	2.88 (10.3)	15.49 (9.99)	-11.44 (10.25)	-10.34 (6.55)	-7.17 (5.06)	-16.54 (7.56)
T	17	16	16	15	14	14
DW	2.02	1.70	2.06	1.77	2.12	1.85
R^2	.977	.962	.983	.906	.90	.922
F	352.0	233.5	338.5	84.2	46.78	128.8

Notes:
* Significant at $\alpha = .01$
** Significant at $\alpha = .05$
*** Significant at $\alpha = .10$

Table 4: Regression Results for Log-Linear Model (Standard Errors in Parentheses)

Independent Variables	Dependent Variables					
	FTRATE	MFTRATE	FFTRATE	PTRATE	MPTRATE	FPTRATE
FTPCI	-.293* (.092)	-.171*** (.089)	-.374* (.102)	-.528** (.183)	-.497** (.181)	-.434** (.175)
RPRVT	.516* (.055)	.472* (.049)	.554* (.068)	.067 (.098)	.041 (.099)	.068 (.097)
UR	.021 (.012)	.019 (.013)	.026*** (.012)	.041*** (.018)	.041** (.017)	.053** (.017)
WD	1.32* (.26)	.797** (.28)	1.86* (.238)	2.68* (.70)	2.29* (.66)	3.22* (.667)
REPJ	-.455 (.276)	-1.02* (.299)	.013 (.266)	.251 (.65)	.217 (.66)	.243 (.67)
CONSTANT	5.97*** (2.89)	11.85* (3.13)	1.0 (2.72)	-2.02 (6.84)	-1.65 (6.89)	-2.44 (6.96)
T	17	16	16	15	14	14
DW	2.07	1.74	2.16	1.75	2.07	1.86
R^2	.984	.972	.988	.908	.905	.927
F	455.7	277.6	472.5	68.6	48.7	111.76

Notes:
* Significant at α =.01
** Significant at α= .05
*** Significant at α = .10

The Impact of Incremental Cost Increases in Successive Monopoly with Downstream Promotion

Peter Brust
Associate Professor of Economics, Department of Finance and Economics
The University of Tampa, E-mail: pbrust@ut.edu

James Fesmire
Dana Professor of Economics, Department of Finance and Economics
The University of Tampa, E-mail: jfesmire@ut.edu

Michael Truscott
Dana Professor of Economics, Department of Finance and Economics
The University of Tampa, E-mail: mtruscott@ut.edu

Abstract:

In a successive monopoly a monopoly manufacturer (upstream firm) sells its product to a monopoly retailer (downstream firm). If there is an increase in the marginal cost of production, the manufacturer will increase the price charged to retailers. Faced with higher cost, the retailer will increase the price charged to consumers, who in turn will purchase a smaller quantity, resulting in a reduction in output and employment for the retailer and a decline in social welfare.

This paper shows that if downstream firms are engaging in standard profit-maximizing behavior, then decreases in output, employment, and welfare resulting from an increase in marginal cost for the manufacturer may be larger than standard economic analysis would suggest. This is because increases in marginal cost may have indirect effects on output and employment by reducing the incentive for retailers to promote their products. When the manufacturer raises the transfer price to the retailer, the marginal returns to promotional activity are reduced. The retailer will reduce the amount of promotion causing a decrease in consumer demand which leads to additional negative impacts on output, employment, and welfare.

I. Introduction

Successive monopoly exists when a monopoly manufacturer (upstream firm) sells its product to a monopoly retailer (downstream firm in the vertical chain of production and distribution). Economic theory shows that increases in the marginal cost of production for manufacturers will lead to increases in the price charged to retailers, who in turn will raise the final price of the product to consumers. As a result, consumers will purchase fewer units of the product, causing a reduction in output and employment for the retailing firm and a decline in social welfare.

This paper suggests that decreases in output, employment, and welfare resulting from an increase in the marginal cost of production for a manufacturer are in fact larger than one would expect from standard economic analysis. This is because increases in marginal cost have indirect effects on employment and welfare by altering the incentives for firms to promote their products. Specifically, downstream firms may react to the producer's transfer price increase by reducing promotional expenditures since the marginal returns to promotion are reduced. This reduction in promotion will decrease the demand for the product, reinforcing the negative impacts on output, employment, and welfare resulting from the increase in the marginal cost of production.

In Section (II) a classic model of successive monopoly is presented and the production-decreasing and welfare-reducing effects are analyzed. Section III introduces downstream promotion to the successive monopoly model, making the point that the incentive to promote a product is determined by the marginal returns to promotion. In Section IV the effects of an increase in the marginal cost of production on the firm's incentive to promote are analyzed. Section V contrasts the effects of an increase in marginal cost on welfare for two downstream firms, one engaging in promotion and the other not. Section VI offers some concluding remarks.

II. Successive Monopoly

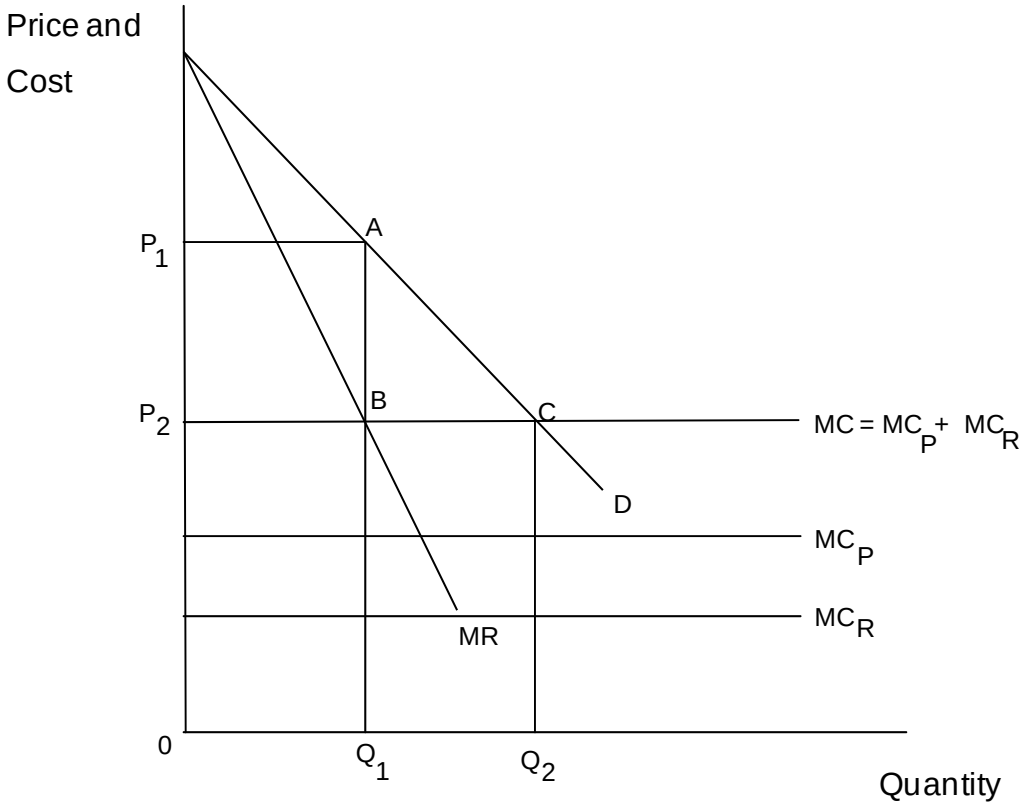
While the successive monopoly model involves monopoly at both stages, it is not necessary for the upstream and downstream firms to be pure monopolists in order for the model to apply. It is only required that both firms have significant market power. Often an upstream firm will grant exclusive distribution territories when there are cost savings associated with having a single distributor in a given geographic area. For example, in newspaper distribution delivery costs are lower with a single distributor servicing all customers on a route rather than having, say, ten agents delivering to every tenth house. "Similarly, the fixed costs of having a display and providing repair service for new automobiles leads the automobile manufacturers to have single distributors in small towns. Multiple dealers would all fail due to large fixed costs." [Blair and Fesmire, 1986, p 62].

The successive monopoly model has been employed frequently in antitrust analysis. Joseph Spengler in his 1950 article entitled "Vertical Integration and Antitrust Policy" [Spengler, 1950] was the first to discuss the arrangement. He argued that the vertical integration of successive monopolies would increase economic efficiency by eliminating the double marginalization associated with independent monopolies at successive stages [Lafontaine and Slade, 2007][Greenhut and Ohta, 1979]. In 1960, Machlup and Taber examined vertical integration in bilateral monopoly and successive monopoly. Blair and Fesmire used the model to analyze the effects of maximum vertical price fixing on the goals of antitrust [Blair and Fesmire, 1986]. In 1990, Fesmire and Romano used the model to examine the results of both maximum and minimum price fixing on social welfare in the presence of downstream promotion [Fesmire and Romano, 1990].

More recently, the model has been employed in analyzing the economic impact of the monopoly relationship between a content provider and a network provider for the online content market [Lanzi and Marzo, 2005]. The content provider's monopoly power derives from its copyrights while the network provider's monopoly power arises from its exclusive ownership of the network loop. The exclusive relationship between Bono Vox and U2.com is a current example of successive monopoly in the online content market. The successive monopoly model was also utilized in the well publicized European Community antitrust case against the merger of AOL and Time Warner. The European Community found that the merger would create a gatekeeper position and dictate standards for one-time musical delivery. In another European antitrust case, the Hildi Case, the successive monopoly model was used to show how a dominant nail gun producer required that its guns use only a specific type of nails. This case involved a dominant nail gun producer requiring that its guns use only a specific type of nails [Zenger, 2005].

Let us begin our analysis of successive monopoly by first examining what happens when both the production and the distribution of a product are controlled by a single monopolist. In Figure 1, D is the demand by consumers for the manufacturer's product, and MR is the associated marginal revenue curve. Let MC_R be the marginal cost of retailing (distributing) the product, and let MC_P be the marginal cost of production. For simplicity, assume both MC_P and MC_R are constant and are therefore equal to average costs. Assume also that transaction costs are zero. Profits are maximized for the monopolist by producing and selling Q_1 units, where marginal revenue (MR) equals the marginal cost of production and retailing ($MR = MC_P+MC_R$). The monopolist will charge the price P_1, resulting in maximum profits equal to $(P_1 – P_2)Q1$.

Figure 1

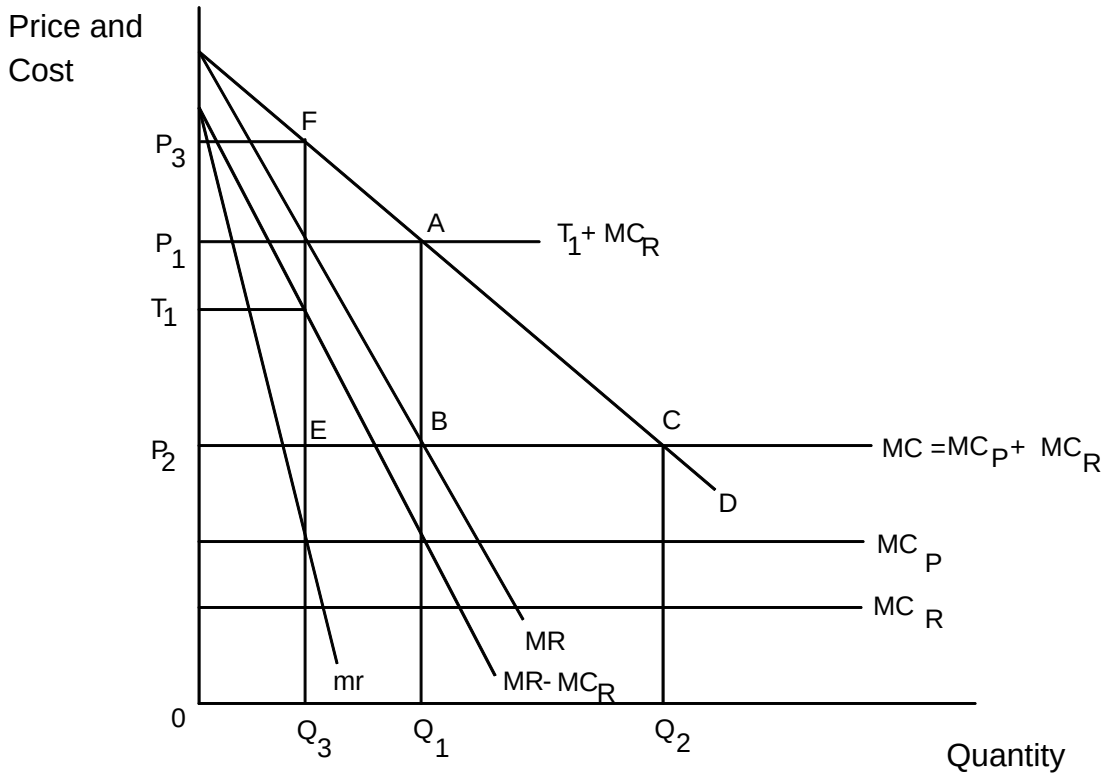


Because the monopolist has the ability to choose price and the resulting quantity, rather than being forced to accept some price dictated by competitive forces, there is a resulting social welfare loss.[1] We can see this by contrasting our monopoly results with those that would occur in a competitive industry. Suppose that we have a competitive industry and assume that each of these small firms is subject to exactly the same production and distribution costs as our monopolist. Competition would drive price to P_2 and quantity to Q_2, where price is equal to average and marginal cost, MC. In Figure 1, area *ABC* represents the loss in social welfare resulting from the monopolist's ability to restrict output. It is equal to the area between the demand curve and the marginal cost curve for the lost output ($Q_2 - Q_1$), which measures the difference between what consumers would have been willing to pay for the lost output and the cost of the resources required to produce that output.

Assume now that a monopoly producer, rather than performing both the production (upstream) and the distribution (downstream) functions, grants exclusive (monopoly) territories to a system of independent distributors to whom he sells the product.

[1] Of course the monopolist is constrained in her choice by the demand curve and its elasticity.

Figure 2



In Figure 2 assume that D, MC_R, and MC_P, are equivalent to those in Figure 1. Since the producer knows that the retailer charges the final price to consumers that will maximize retailer profits, his problem is to choose that unique value of T, the transfer price that the producer charges the retailer, that will induce the retailer to choose that final price and associated value of Q that will maximize the producer's profits. In order to do this, the producer must determine the retailer's demand for his product.

The demand by the retailer for the product produced by the manufacturer is a derived demand. She desires the product only so that she can resell it to consumers. The retailer's demand for the input is derived from the consumer demand for the product.[2] The retailer's marginal cost of supplying an additional unit of the product is equal to $T + MC_R$, the sum of what she must pay the producer and the marginal cost of retailing. Her marginal revenue curve is MR. The retailer chooses that quantity where $MR = T + MC_R$. Rearranging, the result is that the retailer will always choose that quantity where T = MR

[2] Hicks derived rules for the elasticity of derived demand. Derived demand is more elastic the easier it is to substitute other inputs for it, the more elastic is the demand for the final product, the more elastic is the supply of other inputs. See [Hicks, 1932] [Stigler, 1987].

– MC_R. The value of T, on the vertical axis, is related to the quantity purchased by the retailer, on the horizontal axis, by MR - MC_R. Thus, MR – MC_R in Figure 2 is the demand curve for the product faced by the producer and the associated marginal revenue curve is mr.

The producer then maximizes his profits by equating his marginal revenue curve, mr, to his marginal cost, MC_P, therefore selling the quantity Q_3 and charging T_1. This is the transfer price determined by the demand curve of the retailer for the producer's product. Since she must pay the transfer price, T_1, the marginal cost to the retailer becomes $T_1 + MC_R$. The retailer equates her marginal revenue and marginal cost, MR = $T_1 + MC_R$, at Q_3 units and charges P_3. We can now compare the economic results when there is a single monopolist in production and distribution with the results in a successive monopoly situation. The single monopolist maximized his profits by equating his marginal revenue, MR, with his marginal cost, $MC_P + MC_R$, leading him to produce Q_1 units and charge consumers P_1. In the successive monopoly situation, both the upstream and the downstream monopolists have an incentive to restrict output, resulting in a lower quantity for consumers, Q_3, and a higher price, P_3. Recall now that we earlier found the welfare loss due to the output restriction of a single monopolist to be equal to the area ABC. That is, the welfare loss was equal to the difference between the demand curve and the marginal cost curve ($MC_P + MC_R$) for the lost output, $Q_2 – Q_1$. By an analogous argument we see that social welfare loss is greater with successive monopoly, equaling FEC, the area between the demand curve and the marginal cost curve over the greater lost output, $Q_2 – Q_3$. A reorganization of the chain of distribution, replacing a single monopolist with successive monopolists, results in an additional welfare loss equal to the area FEBA.

Paradoxically, even though this reduction in welfare is brought about by an intensified quest for monopoly profits by two monopolists instead of just one, the replacement of a single monopolist with successive monopolists results in decreased total profits. Recall that a single monopoly resulted in to profits equal to the area $(P_1 – P_2)Q_1$. Now that the producer sells Q_3 units to the retailer at a price of T_1 and since the cost per unit to him is MC_P, he makes profits equal to $(T_1 – MC_P)Q_3$ under successive monopoly. The retailer now sells Q_3 units at a price of P_3 and her cost per unit is now $T_1 + MC_R$. Therefore, her profits now are equal to $(P_3 – T_1 – MC_R)Q_3$. The sum of producer profits and retailer profits equals $(P_3 – MC_R – MC_P)Q_3$, or simplified, $(P_3 – P_2)Q_3$**.** By inspection of Figure 2 it is clear that the sum of the profits made by the producer and the retailer under successive monopoly is less than those made by a single monopolist. This provides a powerful motive for the downstream and upstream firms to integrate vertically. If vertical integration is unappealing for some reason, the upstream firm may find it attractive to impose a maximum price on the downstream firm.[3]

[3] The producer could require in Figure 2 that the retailer, as a condition of receiving his exclusive territory, charge a price no higher than P_1, the same price that would maximize profits for a single monopolist. Consumers would purchase Q_1 and the upstream firm would make profits equal to those of a single monopolist. The downstream firm would adopt P_1, equal to $T_1 + MC_R$ and make zero economic profits. But this ignores the possibility of promotional efforts by the downstream firm. See [Fesmire & Romano, 1990] on which our analysis of promotion draws for this paper. Maximum vertical price fixing was illegal until

III. Successive Monopoly and Downstream Promotion

In the classic successive monopoly model of the previous section the upstream firm decides which transfer price to charge the downstream firm, while the downstream firm decides which price to charge the final consumer. Together, these decisions determine the level of upstream and downstream profits. The upstream firm's choice of the transfer price (T) has a significant effect on downstream profits because downstream profits equal $(P - T - MC_R)Q$. The downstream firm's choice of P, which determines Q, has a significant effect on upstream profits, equal to $(T - MC_P)Q$. If the downstream firm engages in product promotion and if there is no contractual agreement on the amount of such promotion, there is a third decision variable, the level of promotional activity the downstream firm chooses. In the absence of vertical price fixing, the downstream firm chooses both the final price and the quantity of promotion. These decisions are closely related because the final price determines, in part, the optimal level of promotion, while the level of promotion, in part, determines the optimal final price.[4]

Advertising and promotion are rich topics which we merely touch upon here. The subjects are covered fully in many texts [Belch and Belch, 2004] [Berman and Evans, 2001]. Concern about the provision of promotional services by a downstream firm goes back at least to Albrecht where the Court said "Maximum prices may be fixed too low for the dealer to furnish services essential to the value which goods have for the consumer or to furnish services and conveniences which consumers desire and for which they are willing to pay" [Albrecht v. Herald Company, 1968, p. 152]. The "services" mentioned by the court are examples of what we call promotion here. Fesmire and Romano examined the effects on output and social welfare when upstream firms impose either maximum or minimum prices on downstream firms when promotion exists downstream [Fesmire and Romano, 1990]. Pauline Ippolito analyzed all 203 reported cases of Resale Price Maintenance (RPM) over the period 1976-1982 [Ippolito, 1988]. She found that the theory that minimum price fixing induces increased promotional efforts by

---

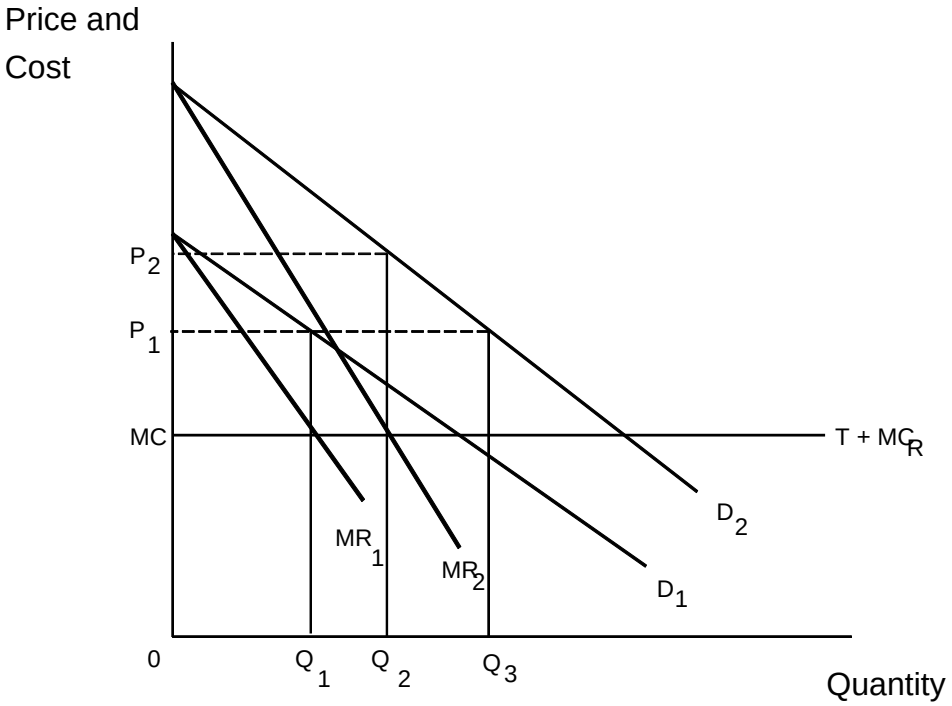
recently when the Supreme Court declared the practice subject to a "rule of reason" approach [Fesmire, 2001]. Quantity forcing is also an option. The firm could insist that the firm sell Q_1 units of the product. To meet this quota the firm would have to charge P_1 and the same result would occur. But this result also ignores downstream promotion and might alienate prospective retailers. Alternatively, the manufacturer could integrate forward with the downstream firm. Vertical integration will be unattractive though if internal transfer costs are greater than market transaction costs [Blair & Kaserman, 1983]. Market transaction costs revolve around two sets of factors whose interaction can increase the costs of market exchange. One set, referred to as "transactional factors", has to do with the difficulties surrounding the negotiation and enforcement of long run contracts because of market uncertainty and the number of potential trading partners available. The second set has to do with "human factors", described as "opportunism" and "bounded rationality" [Williamson, 1974]. Internal transfer costs are not likely to be zero. "This is because 'although the human and transactional factors which impede exchanges between firms (across a market) manifest themselves somewhat differently within the firm, the same set of factors applies to both.'" [Williamson, 1974, pp. 1442-1443] [Blair & Kaserman, 1983, pp. 14].

[4] Promotion as we use it here is a very general concept, including more than advertising. Anything that increases the quantity that some (or all) consumers will purchase at a given price, or the price that some or all consumers will pay for a given quantity, constitutes promotion. Some examples are point-of purchase product demonstrations, product display, the amount/location of shelf space, and anything likely to increase general store traffic.

downstream firms was a potential explanation for approximately fifty percent of the cases of RPM. In addition, she found that maximum price fixing, also involving induced promotional efforts downstream, accounted for twenty percent of her entire sample. This article takes the same induced promotion model developed by Romano and Fesmire and uses it to analyze promotion induced by changes in incremental costs.

We now turn to an examination of downstream promotion. In Figure 3 assume that downstream promotion is zero, the transfer price is fixed, and D_1 is the demand curve. P_1 and Q_1 are the profit-maximizing price and quantity, where $MR_1 = T + MC_R$, the marginal cost of buying and retailing additional units of the product. Suppose now that the downstream firm engages in promotional activity, shifting the demand curve to D_2 and increasing the quantity that consumers want to buy to Q_3 at the existing price, P_1. This increases the firm's revenues by $P_1 \Delta Q$ where $\Delta Q = Q_3 - Q_1$. It also increases the firm's costs by $(T + MC_R)\Delta Q$. The net gain from these promotional efforts to the downstream firm is $(P_1 - T - MC_R)\Delta Q$.

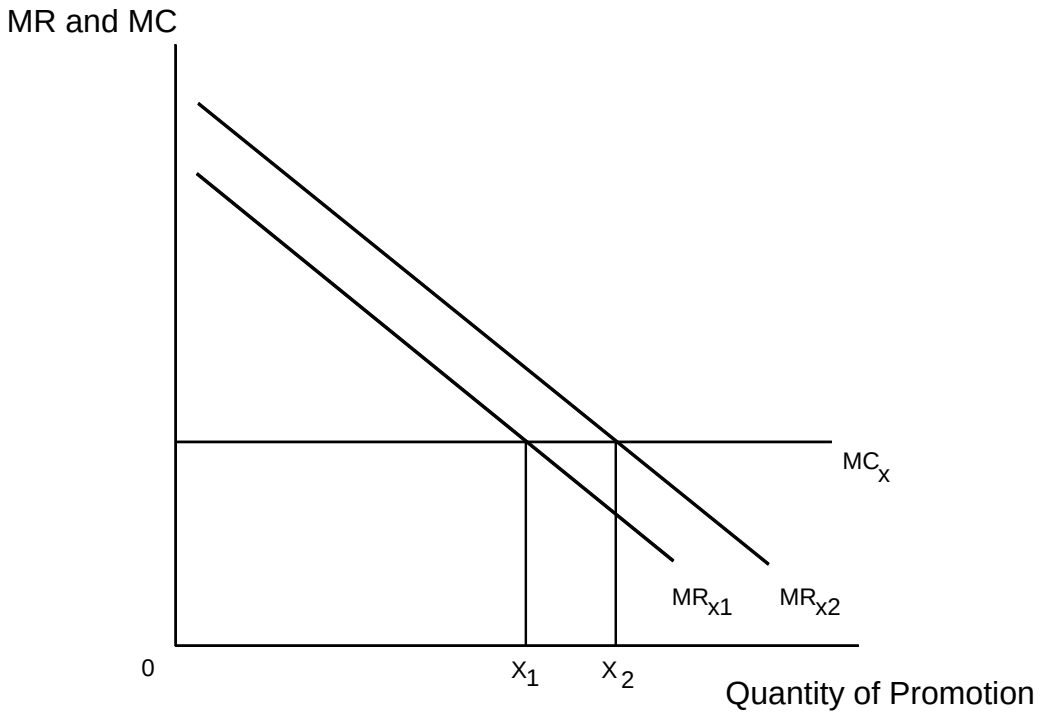
Figure 3



These gains are shown in Figure 4 as MR_{x1}, the net marginal revenue from promotion when the final price is P_1 and where MC_x is the marginal cost of promotional efforts, assumed here to be constant. Since we would expect additional units of

promotion to yield successively smaller increases in quantity when price is constant, MR_{x1} has a negative slope because $(P_1 - T - MC_R)\Delta Q$ decreases as the quantity of promotion increases. The optimal level of promotion is X_1 where $MR_{x1} = MC_x$. Assume the downstream firm engages in this optimal quantity of promotion shifting demand from D_1 to D_2 in Figure 3.

Figure 4

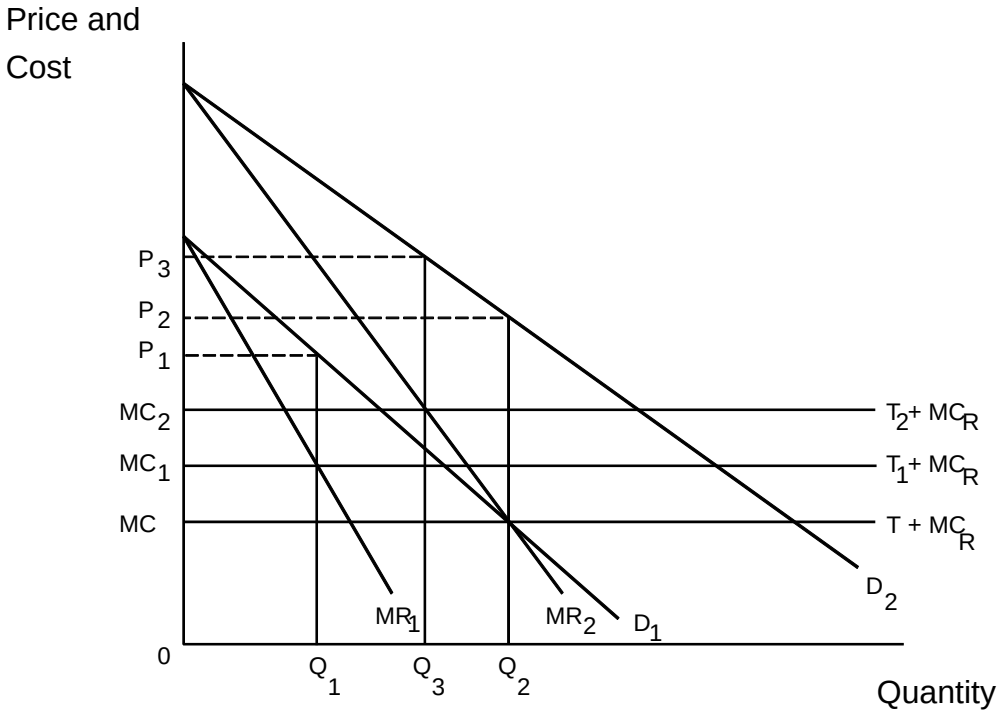


The profit-maximizing price and quantity are now P_2 and Q_2 where $MR_2 = MC = (T+MC_R)$, given that the level of promotion is X_1. When the firm raises its price to P_2, however, the net marginal revenue from promotion in Figure 4 changes to $(P_2 - T - MC_R)\Delta Q$, given by MR_{x2} in Figure 4. The new optimal level of promotion is X_2 where $MR_{x2} = MC_x$.This increase in promotion from X_1 to X_2 would cause a second increase in demand to, say D_3 which is not shown to minimize the clutter in Figure 3. This, in turn, would lead to a new optimal price. Since the optimal level of promotion depends in part on the final price and the optimal price depends in part upon the level of promotion, their levels must be determined simultaneously by the downstream firm. That is, for each possible transfer price the unique price/promotion mix occurs where $MR = (T + MC_R)$ and where $MR_x = MC_x$.

IV. Transfer Prices and the Effect of a Marginal Cost Increase on Promotion

The upstream firm's problem is to choose that value of T that maximizes its profits. The upstream firm knows that if it changes the transfer price, the downstream firm will re-optimize since its incentives will be altered. It is intuitively clear that an increase in the transfer price would cause the downstream firm to choose a new combination of price and promotional effort such that output is lower. This is, of course, because the provision of the final good would be more costly for the downstream firm and also because the marginal returns to promotion, $(P - T - MC_R)\Delta Q$, would be lower for the downstream firm. As the upstream firm raises T, seeking its optimal value, its problem is to balance the effect of an increased profit margin, profit per unit, against the effect of a decrease in the quantity sold, Q. Assume that the upstream firm finds that optimal T so that the downstream firm finds its equilibrium at P_2 and Q_2 in Figure 5 and at X_2 in Figure 4.

Figure 5



Suppose now that marginal cost increases.[5] In Figure 2 this would increase MC_p, the marginal cost of production for the upstream firm. This, in turn, would lead to a new profit-maximizing equilibrium for the upstream firm at a lower quantity than Q_3. Without showing this on the graph, it is cluttered enough, the upstream firm will increase T as discussed above until it achieves a new optimum at a value higher than T_1, the previous optimum. This higher T reduces the marginal returns to promotion, $(P - T - MC_R)$. At the same time the downstream firm adjusts P in that same expression leading to a net change in Figure 4 from MR_{x2} to MR_{x1}. The downstream firm decreases its promotional efforts from X_2 to X_1. In Figure 5 this causes a decrease in demand from D_2 to D_1 and the downstream firm seeks a new equilibrium at P_1 and Q_1 at the intersection of the new marginal revenue curve, MR_1, and the new marginal cost curve, $MC_1 = T_1 + MC_R$.

Now assume for a moment that there is no downstream promotion and that the downstream firm is again in equilibrium at P_2 and Q_2 in Figure 5. Suppose further that the same cost increase mentioned above is imposed. This would increase MC_P in Figure 2 by the same amount mentioned above. The upstream firm will respond by raising T to its new profit-maximizing position. In Figure 5 the downstream firm, faced with an increase in cost, will respond by raising its price to P_3 and consumers will respond on D_2 by reducing quantity demanded to Q_3. This time, however, since there is no downstream promotion, the downstream firm does not cut back on promotion so there is no reduction in demand. With downstream promotion, a given increase in T causes a reduction in Q for two reasons, the higher final price charged by the downstream firm and the reduction in demand caused by the reduction in promotion. With no downstream promotion, a given increase in T causes a reduction in quantity for only one reason, the higher price charged by the downstream firm.[6]

Since a given increase in T causes a smaller decrease in its sales, the downstream firm will increase the transfer price more when there are no downstream promotional effects than if these effects were present. Assume the downstream firm increases transfer price. This increases marginal cost to $MC = T_2 + MC_R$ in Figure 5 and the downstream firm maximizes profits at P_3 and Q_3. When promotional effects are present, an increase in incremental costs for the upstream firm leads to a greater reduction in output. We next turn to a comparison of the welfare effects of a marginal cost increase with downstream promotion and without downstream promotion.

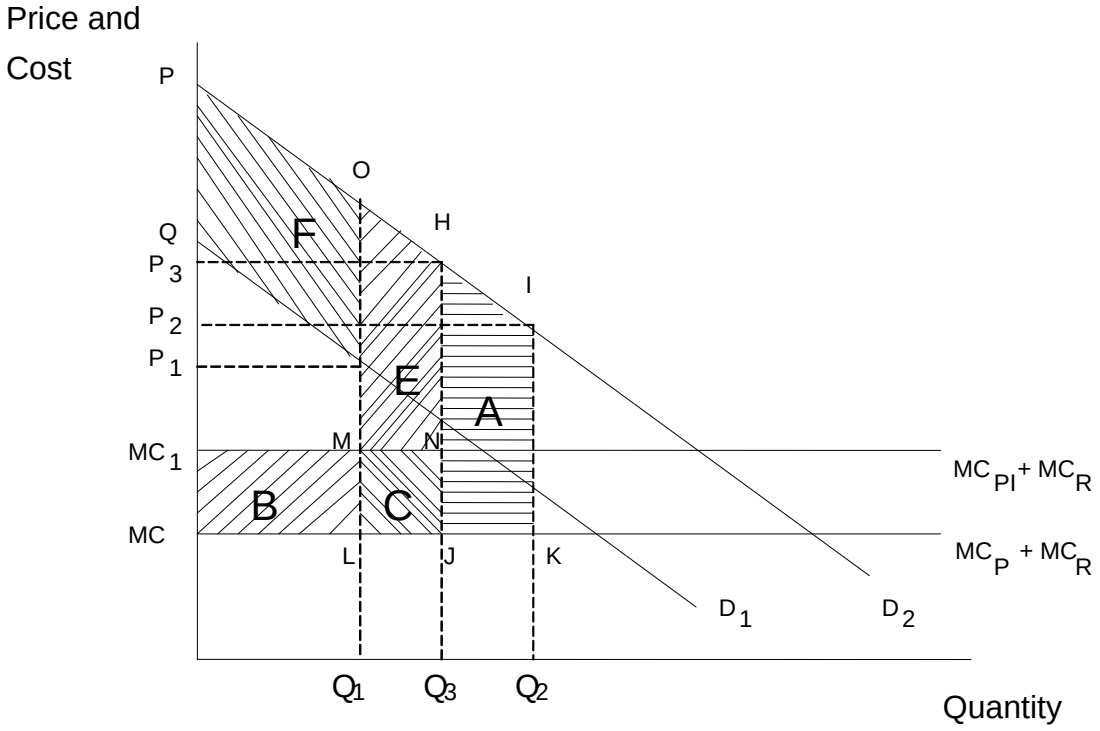
[5] Such a cost increase may emanate from a number of sources including: 1) a decrease in the value of the dollar that causes the price of imported inputs to increase: 2) new environmental regulations such as a carbon tax or a "cap and trade" system to control carbon dioxide emissions; 3) increases in wages or government regulations.

[6] We should point out here that we assume that the downstream firm engages in classic profit maximization by producing where MR = MC. This is often not the case because downstream firms sometimes do not know enough about their demand. Because of this retailers often engage in other forms of pricing. In fact, if the downstream firm engages in cost plus pricing, an increase in marginal cost for the upstream firm may very well result in an "increase" in promotional efforts by the downstream firm.

V. The Welfare Effects Compared

Figure 6 duplicates the demand and cost curves shown in Figure 5. Assume we have two monopoly downstream firms, one engaging in downstream promotion, the other not. Both downstream firms begin at P_2 and Q_2 on D_2. For simplicity, assume that both firms are served by the same upstream monopolist and are therefore participants in successive monopoly. $MC = MC_P + MC_R$ is the marginal cost curve for both successive monopolies. Assume that MC_P increases so that the marginal cost of production and retailing increases to $MC_1 = MC_{P1} + MC_R$ for both firms. As a result, the downstream firm without promotion moves to P_3 and Q_3 on the existing demand curve, D_2. The firm with promotional effects moves to P_1 and Q_1 on its new demand curve, D_1, which has decreased because of the reduction in promotion mentioned above.

Figure 6



Both successive monopolies suffer a loss of sales from Q_2 to Q_3. This results in a welfare loss of shaded area A, equal to JHIK. There is a welfare loss equal to B, $MCMC_1ML$, in both markets as the cost of providing units from zero out to Q_1 increases for both firms from $MC_P + MC_R$ to $MC_{P1} + MC_R$. The firm without promotion continues to sell units from Q_1 to Q_3, but there is a welfare loss equal to area C, LMNJ, because those units now cost more. The firm with promotion no longer sells units Q_1 to Q_3. This results in a welfare loss equal to C + E, LOHJ. In addition, for the firm with promotion there is a decrease in welfare for units from 0 to Q_1 measured by the difference between D_2, the value to consumers before the decrease in promotion, and D_1, reflecting the good's new value to consumers, area F.[7]

In sum, the welfare loss associated with an increase in marginal cost when there is downstream promotion is equal to A + B + C + E + F, while the welfare loss with no promotion is equal to A + B + C. The actual difference would be somewhat reduced if the reduction in promotion costs for the firm with downstream promotion are taken into account. These costs will often be approximately offset by the decrease in consumers' inframarginal valuations.[8] The magnitude of these changes in welfare will vary directly with the magnitude of the changes in T and with the degree of monopoly power possessed by the successive monopoly.

VI. Conclusion

Firms are constantly reacting to changing market conditions by altering their price and output strategies. Predictably, increases in incremental costs would result in price increases and output reductions with an accompanying decrease in welfare. This paper shows that these negative effects may be larger than recognized by basic economic theory if the downstream firms alter their promotional strategy in response to cost increases. The magnitude of these effects may vary significantly among different industries which opens the opportunity for empirical research in this area. As upstream firms increase transfer price output should decline more in industries with downstream promotion than in those without promotion. In the area of public policy, the theoretical framework of this paper, along with supportive research, should aid lawmakers in more accurately evaluating the consequences of policy decisions which affect business costs.

[7] Our assumption here is that demand accurately reflects consumers' valuations. There is a social benefit to promotion, for example by providing information to the uninformed. There is no deception in promotion [Mathewson & Winter, 1984]. There are other approaches. Dixit and Norman assume there is a single, unalterable marginal valuation function, which may or may not coincide with advertising-influenced demand [Dixit & Norman, 1978].

[8] This will be so if the decreases in promotion cause parallel shifts downward in demand and if the changes are not too large. See [Boudreaux & Ekelund, 1988] where they argue that vertical shifts are the only reasonable result of promotion. [Scherer, 1984] would object to a focus on the case of parallel shifts. He emphasizes the importance of cases where shifts are otherwise. His analysis depends on downstream competition, which is not present here.

References

Albrect v. Herald Co. (1968). 390 U.S. 145.

Belch, G.E. and M.A. Belch (2004). *Advertising and Promotion: An Integrated Marketing Communications Perspective*, Boston: McGraw-Hill/Irwin.

Blair, R. D. and D. L. Kaserman (1983). *Law and Economics of Vertical Integration and Control*. New York: Academic Press.

Blair, R. D. and J. M. Fesmire (1986). "Maximum Price Fixing and the Goals of Antitrust," *Syracuse Law Review*, 37 (1): 43 - 77.

Boudreaux, D. J. and R. B. Ekelund Jr. (1988). "Inframarginal Consumers and the Per Se Legality of Vertical Restraints," *Hofstra Law Review*, 17: 137 - 158.

Dixit, A. K. and V. Norman (1978). "Advertising and Welfare," *Bell Journal of Economics*, 9 (1): 1 - 17.

Fesmire, J. M. and R. E. Romano (1990). "Maximum Versus Minimum Price Fixing and Promotional Incentives: Economics, Law, and Policy," *George Mason University Law Review*, 13 (2): 263 - 301.

Fesmire, J. M. (2001). "Maximum Vertical Price Fixing From *Albrecht* Through *Brunswick* to *Khan*: An Antitrust Odyssey," *Seattle University Law Review*, 24 (3): 721 - 762.

Greenhut, M. L. and H. Ohta (1979). "Vertical Integration of Successive Oligopolists," *American Economic Review*, 69 (1): 137 - 141.

Hicks, J. R. (1932). *The Theory of Wages*. London: Macmillan.

Ippolito, P. M. (1988). "Resale Price Maintenance: Economic Evidence from Litigation," Washington, D. C.: Federal Trade Commission.

Lafontaine, F. and M. Slade (2007). "Vertical Integration and Firm Boundaries: The Evidence," *Journal of Economic Literature,* 45 (3): 629 - 685.

Lanzi, D. and M. Marzo (2005). "Content Delivery and Vertical Integration in On-line Content Markets," *Review of Network Economics*, 4 (1) March: 63 - 74.

Machlup, F. and M. Taber (1960). "Bilateral Monopoly, Successive Monopoly, and Vertical Integration," *Economica: New Series*, 27 (106): 101 - 119.

Mathewson, G. F. and R. A. Winter (1984). "An Economic Theory of Vertical Restraints," *RAND Journal of Economics*, 15 (1): 27 - 38.

Scherer, F. M. (1984). "The Economics of Vertical Restraints," *Antitrust Policy in Transition: The Convergence of Law and Economics*, E. Fox and J. Halverson eds., Chicago: American Bar Association.

Spengler, J. J. (1950). "Vertical Integration and Antitrust Policy," *Journal of Political Economy,* 58 (4): 347 - 352.

Stigler, G. J. (1987). *The Theory of Price*, New York: Macmillan.

Williamson, O. E. (1979). "The Economics of Antitrust: Transaction Cost Considerations," *University of Pennsylvania Law Review*, 122: 1442 - 1443.

Zenger, H. (2005). "Successive Monopolies with Endogenous Quality," Munich, University of Munich, zenger@lmu.de: 1 - 29.

Promotional Payments and Firm Characteristics: A Cross-Industry Study

Adam D. Rennhoff
Assistant Professor of Economics, Department of Economics and Finance
Middle Tennessee State University, E-mail: rennhoff@mtsu.edu

Abstract:

This research uses publicly available financial data to conduct a cross-industry study of the prevalence and magnitude of promotional payments. The passage of new accounting regulations provides a unique opportunity for a closer inspection of the practice of promotional payments. In this paper, I develop a simple theoretical model of retailer-manufacturer bargaining. The model predicts that firms with high price-cost mark-ups and market share should make higher promotional payments. Industries with less concentration should be associated with higher promotional payments, as well. I test these predictions using a novel new dataset. I find empirical support for the predictions made with the theoretical model.

I. Introduction

Manufacturers often pay retailers in exchange for a benefit at the retail outlet, such as premium shelf space or an end-of-aisle display. The details of these promotional payments have been well-guarded secrets for many years. Agreements regarding the dollar amount paid or other considerations were often off-invoice. Manufacturers and retailers, alike, have argued that this secrecy is necessary in order to protect confidential business strategies. Coca-Cola, for example, has no interest in allowing Pepsi to gather information regarding Coca-Cola's dealings with retailers. Similarly, Albertsons would prefer to keep their (incoming) payments and promotional strategies secret from Kroger.

In November 2001, the Financial Accounting Standards Board's (FASB) Emerging Issues Task Force (EITF) introduced EITF Issue No. 01-9, "Accounting for Consideration Given by a Vendor to a Customer or a Reseller of the Vendor's Products."[1] These new accounting regulations have fundamentally changed the way companies account for promotional allowances. Prior to the rule changes, there were no stated requirements regarding where (in their 10-K SEC filings) firms were to account for any promotional payments made. While the majority considered promotional payments to be an expense and included it in the "Selling, General, and Administrative expenses" category, there was no uniform approach. Note that in financial statements, unless directed otherwise, most information is presented at an aggregate level. As such, the

[1] Issue 01-9 consolidated and codified two previous EITF issues: Issue No. 00-25, "Accounting for Consideration from a Vendor to a Retailer in Connection with the Purchase or Promotion of the Vendor's Products," and Issue No. 00-14, "Accounting for Certain Sales Incentives." EITF Issue No. 00-25 addressed the accounting treatment and classification of various types of "consideration" given by manufacturers, such as buy-downs and slotting fees. Issue No. 00-14 provided guidance on the accounting treatment of sales incentives aimed at consumers, such as manufacturer coupons.

amount spent on promotional payments was "hidden" in all years prior to the change in accounting standards.

EITF Issue No. 01-9 formalized the way in which firms should account for promotional payments. Instead of allowing firms to account for payments as they saw fit, the rules specified that all promotional payments should be considered as a reduction in net revenues. Moreover, this change was retroactive, requiring firms to go back (at least) one year to restate their prior net revenue (taking promotional payments as a reduction in revenues). The fact that firms needed to go back and restate net revenue forced them to explicitly state the amount of promotional payments made in the previous year.[2] This event gives us a one year window through which we may investigate promotional allowances.

While it would be ideal to create a panel dataset, looking at the use of promotional allowances over time (for a number of firms), that is simply not feasible. In the years since the accounting rule change, firms were required to continue the practice of treating promotional payments as a reduction in net revenue. However, they were not required to reveal the amount of promotional payments made each year (and so, not surprisingly, none of them did). They are only required to report the aggregate revenue total, not its components. Based on the fact that financial reports are not truly transparent, the opportunity to observe exactly how much a given firm allocated towards promotional payments was a one-time opportunity, attributable solely to the requirement that firms restate revenue from the preceding year.

Recent events, such as regulatory inquiries into A&P Supermarkets' and K-Mart's mishandling of "vendor allowances", have brought renewed interest in examining an issue that has been around for many years. For example, Fortune magazine published an article on retail trade promotions "careening out of control" in their July 1983 issue. According to this 1983 article, spending on promotional allowances had grown from $1 billion (annually) in the early 1970's to roughly $8 billion at the time of publication. This value figures to be even higher today. In particular, more recent studies on slotting allowances, which may be considered as a subset of general promotional allowances, have estimated spending at $16 billion annually [Desiraju, 2001].

Given the dollar amounts spent annually on promotional allowances, it is surprising how little we truly know about the practice. This fact is due, in large part, to the lack of publicly available data. Fortunately, the FASB accounting rule changes offers a rare glimpse into the practice. The primary purpose of this paper is to make use of this publicly available accounting data to conduct an empirical study on promotional allowance payments. The data allows me to test the predictions from a simple model of manufacturer-retailer negotiation. This bargaining model produces predictions regarding which firm and market characteristics are likely to be related to the practice of paying promotional allowances. The empirical model used to test the theoretical predictions is estimated using a two-stage procedure that accounts for the interrelatedness of a firm's

[2] The restatement in net revenue was typically mentioned in the "Footnotes" section of a firm's 10-K Annual Report.

decision of whether to offer promotional allowances and, conditional upon that decision, the choice of how much to spend on promotional allowances. Results from the empirical model provide statistically significant confirmation of the theoretical predictions.

To my knowledge, this paper represents the first attempt to use publicly available financial data to conduct a cross-industry study on the prevalence and magnitude of promotional payments. I now turn to a description of the theoretical model.

II. Manufacturer-Retailer Bargaining

In this section, I propose a simple model of manufacturer-retailer bargaining. This model will be used to generate predictions regarding the use (and magnitude) of merchandising allowance payments.[3] In the game, a retailer is bargaining with a manufacturer over two decision variables: the per-unit wholesale price (w) paid by the retailer and a lump sum promotional allowance payment (A) paid to the retailer.

A. Disagreement Profits

In the event that an agreement cannot be reached, each of the parties receives their respective disagreement profit. I define Π_0^R as the retailer's disagreement profit. In the case of an upstream monopoly, we may think of Π_0^R as being equal to zero. If, on the other hand, the retailer carries numerous other products (possibly in other product categories), then the retailer's disagreement profit should be strictly greater than zero. In this manner, I assume that Π_0^R is greater than or equal to zero.

I similarly define Π_0^m as the manufacturer's disagreement profit. If there is a downstream monopoly, then this is likely to be zero. If, on the other hand, the manufacturer has multiple selling outlets, then Π_0^m would be positive. Indeed, we might interpret larger values for Π_0^m to indicate less dependence on a particular retailer.

B. Agreement Profits

In the event that an agreement is reached, both parties realize profits, which I denote Π^R and Π^m. These profits resemble traditional profit equations. These agreement profits can be written:

$$\Pi^R = (p - w)Q(p) + A \tag{1}$$

$$\Pi^m = (w - c)Q(p) - A \tag{2}$$

[3] Ideally, predictions regarding these factors would come from previous (detailed) theoretical work. Unfortunately, theoretical research on promotional allowances tends to focus solely on slotting allowances. The term "slotting allowance" usually refers to payments made in order to induce a retailer to carry a new product, while "promotional" or "merchandising allowances" often encompass a wider variety of acts, such as in-store display or advertisement. It is not clear, therefore, that conclusions made regarding slotting allowances are necessarily applicable to promotional allowances. For a survey of the slotting allowance literature, see Bloom, et al. [2000].

where p is the per-unit retail price of the good, Q are the sales (which is a function of p), c is the marginal cost of production, and w and A are as defined above. In order to make the problem tractable, it is necessary to impose some assumptions regarding functional forms. For simplicity, I assume that downstream demand can be represented by a linear function and that the retailer's reaction function ($p(w)$), which is the optimal retail price expressed as a function of the upstream wholesale price, is proportional to the manufacturer's wholesale price. Specifically:

$$Q(p) = \alpha - \beta p$$
$$p(w) = \delta w$$

where $\beta > 1$ and $\delta \geq 1$.[4] This allows the agreement profits to be written:

$$\Pi^R = \left((\delta - 1)w\right)\left(\alpha - \beta\delta w\right) + A \quad (3)$$
$$\Pi^m = \left(w - c\right)\left(\alpha - \beta\delta w\right) - A \quad (4)$$

The Nash bargaining equilibrium is the solution to the following problem:

$$\max_{w,A} \left(\Pi^R - \Pi_0^R\right)^\theta \left(\Pi^m - \Pi_0^m\right)^{1-\theta} \quad (5)$$

such that $A \geq 0$, $w \geq 0$, $\Pi^R - \Pi_0^R \geq 0$, and $\Pi^m - \Pi_0^m \geq 0$. These first two constraints may be called "reality" constraints and the latter two are incentive compatibility constraints. I am interested in promotional allowances, which flow from manufacturer to retailer. I, thereby, eliminate the possibility that the retailer, instead, pays these allowances to the upstream firm (i.e. $A < 0$). Similarly, I do not allow for the possibility that the manufacturer pays the retailer for each unit the retailer buys (i.e. $w < 0$). In the maximization problem above θ (1 - θ) is retailer's (manufacturer's) bargaining "power."

C. Solutions and Comparative Statics

Case #1 (Disagreement Profits are Zero): In the case where disagreement profits and the marginal cost are assumed to be zero, the solutions to the Nash bargaining problem are:

$$w^* = \frac{\alpha}{2\beta\delta}$$
$$A^* = \frac{\alpha^2\left(1 - \delta\left(1 - \theta\right)\right)}{4\beta\delta}$$

[4] The latter condition simply implies that the retailer does not price below cost. In this specification, δ-1 is the retailer's mark-up percentage.

Notice that A^* is not unambiguously greater than zero. A^* will be greater than zero if and only if $\delta < \frac{1}{1-\theta}$. This says that the allowance will be greater than zero as long as the retailer's mark-up is not "too high." The primary purpose of the empirical sections below is to determine how the value of A^* changes with differences in firms and industries. As such, there are a number of comparative statics that might be of interest. Below are several of these comparative statics and a brief interpretation of what this finding may mean for the empirical model.

- $\frac{\partial A^*}{\partial \theta} = \frac{\alpha^2}{4\beta} > 0 \Rightarrow$ As the retailer's bargaining power increases (decreases), the agreed upon allowance payment should be higher (lower)
- $\frac{\partial A^*}{\partial \alpha} = \frac{\alpha(1-\delta(1-\theta))}{2\beta\delta} \gtreqless 0 \Rightarrow$ This comparative static is ambiguous. However, as long as $A^* > 0$, then $\frac{\partial A^*}{\partial \alpha} > 0$, which indicates that demand shifts lead to higher allowances

$\frac{\partial A^*}{\partial w^*} > 0 \Rightarrow$ As the wholesale price increases, the allowance increases as well.[5]

This last comparative static can be seen by inserting w^* into A^*, which yields $A^* = \frac{\alpha w^*(1-\delta(1-\theta))}{2}$. The derivative of which, with respect to w^*, is positive as long as $A^* > 0$.

Case #2 (Non-Zero Disagreement Profits): In the case where disagreement profits are no-longer equal to zero, the solutions to the Nash bargaining problem are:

$$w^* = \frac{\alpha}{2\beta\delta}$$

$$A^* = \frac{\alpha^2(1+\theta\delta-\delta)+4\beta\delta\Pi_0^R-4\theta\beta\delta\left(\Pi_0^R+\Pi_0^m\right)}{4\beta\delta}$$

In addition to the comparative statics presented in the previous section, utilizing non-zero disagreement profits allows me to examine several other interesting comparative statics.

- $\frac{\partial A^*}{\partial \Pi_0^m} = -\theta < 0 \Rightarrow$ As the value of the manufacturer's outside options increase, the offered allowance decreases

[5] This holds only when $A^* > 0$.

- $\frac{\partial A^*}{\partial \Pi_0^R} = 1 - \theta > 0 \Rightarrow$ As the value of the retailer's outside option increases, the agreed upon allowance will also increase

It can easily be shown that the comparative statics from the preceding section, such as $\frac{\partial A^*}{\partial \delta} < 0$, also hold in the non-zero disagreement profits specification. The solutions and comparative statics give me a number of testable implications, which I outline in the section that follows.

D. Testable Implications

In this section, I describe three testable implications from the theoretical model. All three involve predictions regarding how the magnitude of promotional allowances are likely to be related to firm and industry characteristics.

Hypothesis 1. *Manufacturers with higher mark-ups (over cost) and market share should also make larger allowance payments*.

This hypothesis is related to the finding that $\frac{\partial A^*}{\partial w^*} > 0$. If we assume that marginal costs of production are fixed, then a manufacturer with a higher wholesale price-cost margin (which implies a higher w) should pay larger promotional allowances. Note that this hypothesis is also consistent with an anti-competitive story where "strong firms" (so denoted by their ability to increase their mark-up) also make larger allowance payments, which effectively bids up the price of in-store promotion for all other manufacturers. Because a firm's ability to increase their mark-up is often related to their market share, we may also expect larger allowances to be paid by firms with higher market share. To see this relationship, note that the manufacturer's equilibrium mark-up over marginal cost $(w^* - c)$ can be written: $(w^* - c) = \frac{1}{\beta\delta} Q^* - c$, where c is a constant marginal cost (which has been assumed to be zero in the previous analysis) and Q^* is the manufacturer's equilibrium level of sales. The derivative of this mark-up is positive, indicating that higher sales (and, therefore, higher market share) imply higher manufacturer mark-up.[6] The higher S_j^*, the larger the implied mark-up. So, from Hypothesis 1, there are two testable predictions: that mark-up, which will be proxied in the empirical model by a firm's gross margin (see Table 3 for a definition), and market share are both positively related to the amount of promotional allowances paid.

Hypothesis 2. *Industries with less upstream concentration are characterized by larger promotional allowance payments*.

[6] This positive relationship between market share and mark-up also holds in the logit model of demand. A derivation of this can be seen in a number of papers, such as Anderson and de Palma [2001].

This hypothesis is related to the disagreement profits of the retailer. The greater the upstream competition (as measured through a less concentrated upstream industry), the stronger the position of the downstream retailer. In the extreme, consider the case of perfectly competitive manufacturers and a monopolist retailer. This retailer "strength" can be measured either as a high θ or a high Π_0^R, both of which are positively related to the size of promotional allowance payments (in the theoretical model). Empirically, I use an industry's Herfindahl Index as a measure of the degree of upstream concentration. The testable implication of Hypothesis 2, therefore, is that an industry's Herfindahl should be negatively related to the amount of promotional allowances paid. Table 1 summarizes the testable predictions.

TABLE 1
Testable Implications

	Prediction	Empirical Test
Hypothesis #1	A firm's mark-up is positively related to the amount of promo. allowances paid	Gross-margin > 0
	A firm's market share is positively related to the amount of promo. allowances paid	Market Share > 0
Hypothesis #2	Industries with less upstream concentration should be characterized by larger allowance payments	Herfindahl < 0

As these hypotheses clearly indicate, a manufacturer's propensity to pay promotional allowances to a retailer depend not only on characteristics of the manufacturer, but also on characteristics of the industry, itself. Using a unique data set, collected specifically for this research, I control for industry differences and examine how various firm-specific factors are related to allowance payments.

III. Data & Variable Measurement

The data on promotional allowance payments used in this study come from corporate Annual Reports (Form 10-K) for 2001. Careful consideration was given to selecting the firms used in this study. I began by compiling an extensive list of manufacturing firms using lists, such as those compiled by the Fortune 1000 and the Grocery Manufacturers of America. The data set includes only those manufacturers that produce pre-packaged consumer goods, sold through retailers that carry brands from multiple manufacturers. This qualification is established to eliminate vertically integrated firms, such as the Gap, as well as direct-to-consumer sellers, such as Dell Computers. Based on these qualifications, the initial sample included 252 manufacturers.

For each of the firms in the sample, information regarding the amount of "consideration" a manufacturer paid can be found in the "Footnotes" section of the firm's 10-K filing. There is great variation in the level of detail each firm presents. For example, the Monterey Pasta Company breaks down their spending on Issue 01-09

related consideration, into distinct categories ("slotting fees and promotions", for example), while Pepsico reports only the total amount. To maintain consistency across the varied reporting techniques, only the total amounts reclassified (due to EITF Issue 01-09) are included in the data.

A number of the firms initially selected were, ultimately, excluded from the study due to one of the following three common reasons: (1) ambiguously worded annual reports that made it difficult to ascertain whether payments were made (14 firms), (2) the firm had not yet adopted the accounting standards requiring disclosure of consideration payments (57 firms), and (3) foreign-owned or foreign-listed firms that adhere to different accounting standards (10 firms). In total, the final sample consists of 171 firms, 100 firms that have been identified as making promotional payments and 71 that do not. Summary statistics appear in Table 2.

TABLE 2
Summary Statistics

	Firms with Payments	Firms without Payments	All Firms
n	100	71	171
Average Payment ($ millions)	254.7	N/A	144.4
Average Gross Margin	0.42	0.36	0.39
Average Inventory Turnover	337.04	341.64	339.03
Average Sales ($ thousands)	5555.32	7042.53	6199.38
Average Market Share (%)	9.2	11.5	10.2
Average Herfindahl	2549	2999	2744
Average Return on Equity (ROE)	14.02	3.28	9.56
Grocery (#)	51	12	63
Software (#)	6	9	15
Home (#)	10	9	19
Electronics (#)	7	14	21
Media (#)	9	9	18
Pharm (#)	5	2	7
Person (#)	8	1	9
Clothes (#)	4	5	9

The firm and industry characteristics listed in the summary statistics are defined in Table 3 below:

TABLE 3
Explanatory Variables

Variable	Definition
Marg	A firm's gross margin (total revenue minus the cost of goods produced, divided by sales)
Inv	A firm's inventory turnover (cost of goods sold (COGS) divided by inventory)
Share	A firm's share of total sales within their industry category
Herf	Industry Herfindahl index calculated using the firm shares
ROE	A firm's return on equity (ROE)

The three hypothesis-testing variables are gross margin (hypothesis 1), market share (hypothesis 1), and industry Herfindahl Indices (hypothesis 2). Since a manufacturer's gross margin captures the relationship between total revenue and cost as a percentage of sales, it can be used as a proxy for their mark-up.

The calculation of manufacturer market share and industry concentration (as measured through the Herfindahl Index) rely on SEC Standard Industrial Classifications (SIC). Each of the 171 manufacturers in our sample is grouped, using the SIC as a guideline, into one of the following broad categories:

1.Grocery products (food and beverage)
2.Computer Software
3.Home products (hardware, tools, furniture)
4.Electronics (computers, stereos)
5.Media (newspapers, magazines)
6.Pharmaceutical products (including over-the-counter drugs)
7.Personal care products (make-up, deodorants, health and beauty)
8.Clothing/Apparel
9.Other

The first eight industry categories are relatively straightforward and can easily be derived from the SIC guidelines. The "Other" listing is included to account for several manufacturers for which it is difficult to place them in any one industry category.

Using manufacturer-level sales figures, I am able to derive each firm's percentage of total industry sales (market share) and, subsequently, each industry's measure of concentration (Herfindahl). Each firm within a category will have the same Herfindahl, but the values will vary across industries.

In order to better explain the firms' promotional decisions, I include two other variables as controls: a firm's inventory turnover and their return on equity (ROE). The inventory turnover variable is included to help address how well each manufacturer's product "moves." A high inventory turnover value might indicate that the manufacturer is a strong seller, thereby making them less risky to the retailer. A priori, one might expect brands with high inventory turnover (and, therefore, lower risk) to be less likely to make large payments to retailers, although this is not explicitly captured in the theoretical model.[7]

ROE is included to examine the (potential) relationship between a firm's profitability (or at least their perceived profitability or value) and their promotional allowance strategy. The empirical model is not sensitive to my choice of perceived profitability; other possible measures, such as price-earnings ratio (P/E), market capitalization, and stock price, yield virtually identical results.

The sample selection criteria eliminated 57 firms that had not yet adopted the new accounting standards.[8] The fact that some firms had not yet adopted the accounting standards raises the concern that the firms included in my sample are not necessarily representative of the general population. If there is some commonality among the firms adopting the standards "on time," then the potential for bias may be present. To address this potential source of bias, I estimated a binary probit model using 228 firms.[9] The dependent variable in this model is a dummy indicating whether a firm had adopted the accounting standards. The purpose of this exercise is to see whether any observable firm characteristics are correlated with a firm's adoption decision. The results of the binary probit, which are presented in Table 4, give me confidence that selection bias is not problematic in this instance.

[7] This presumption is based on the argument that retailers request allowances, in part, to either safeguard against poor sales or cover some of the cost of storing a product.

[8] When mentioning the accounting rule change in the "Footnotes" of their 10-K, most of the firms that had not yet adopted the standard stated something similar to the following: "We have not yet adopted the practices outlined in EITF No. 01-9. At this time, we are studying the impact this will have on our statement."

[9] The 228 firms include the 171 firms that adopted the accounting standard plus the 57 firms that had not. I continue to exclude the 24 firms that either had ambiguous reports or were foreign-owned.

TABLE 4
Probability of FASB Accounting Standards Adoption
Probit Marginal Effects (Adopt =1)

	I	II
Margin	-0.6130	-0.5701
	(2.0124)	(1.1627)
Inv	0.0000	0.0000
	(0.0001)	(0.0001)
Share	-0.3300	-0.1692
	(0.3070)	(0.4112)
Herf	0.0001	0.0002
	(0.0004)	(0.0076)
ROE	0.0021	0.0025
	(0.0011)	(0.0027)
Log Likelihood	-107.52	-106.28
Industry Dummies	No	Yes

* -- Significant at the 1% level
a -- Significant at the 5% level

In the specification with no industry dummies, there are no statistically significant explanatory variables. When industry dummies are included, the only statistically significant variable (significant at the 10 percent level) is the dummy variable associated with the grocery industry. All coefficients of interest remain insignificant. I conclude, therefore, that there is no systematic bias due to the sample selection criteria.

IV. The Estimation Model

The predictions discussed in Table 1 refer to the factors that influence how much each firm pays in promotional allowances (i.e. the magnitude of their payments). However, to properly estimate a model of this decision, I must also account for a related decision: whether to pay promotional allowances at all. These two decisions are interrelated because unobserved characteristics that influence a firm's decision to offer promotional allowances may also influence the decision regarding the amount of promotional allowance spending. To properly account for this, I adopt the modeling framework of Krishnamurthi and Raj [1988] and Nagler [2006] and estimate a two-stage model in which I first estimate the decision of whether to offer promotional allowances and then estimate the amount of promotional allowances paid (using the subsample of firms that pay promotional allowances), incorporating a "conditionality term."

More specifically, the binary offer-allowances-or-not (first-stage) decision of manufacturer *i* (in industry *m*) can be written:

$$\Pr\left(y_i = 1\right) = \Pr\left(y_i^* > 0\right)$$

where

$$y_i^* = \beta' x_i + \alpha Z_m + \varepsilon_{im}$$

and where y_i^* is a latent variable, which is unobservable to the econometrician, x_i is a vector of observable firm characteristics, Z_m is a vector of industry-specific variables (dummies), and ε_{im} is a mean-zero idiosyncratic error term.

The second stage model examines how the amount of promotional allowance payments are related to firm and industry characteristics. This second stage will be used to compare empirical reality with the predictions from the theoretical model. The dollar amount of allowances paid by manufacturer i can be written:

$$y_i = \theta' x_i + \gamma Z_m + \delta S_i + \eta_i \quad \text{for } y_i > 0$$

where y_i is the dollar amount of promotional allowance spending by firm *i*, x_i and Z_m are as defined above, S_i is the conditionality term, and η_i is the error term. The conditionality term (S_i) is the fitted probability that firm *i* offers any promotional allowances. This probability is obtained using the estimates from the first stage.

The first stage is estimated using a binomial probit (therefore, ε is distributed standard normal). The second stage is estimated as a Tobit, with an assumed lower bound on the dependent variable (y_i) of zero.

V. Results

The estimation results for the manufacturer's two decisions appear in Tables 5 and 6. In each table, I present results with and without industry dummy variables. Given the likelihood that there are unobserved industry characteristics, the use of industry dummies is preferable. It should be noted that the inclusion of these dummy variables does not qualitatively change the results. The predictions from the second stage model are of primary importance, as they test the predictions set out in the theoretical model, so I discuss them first.

TABLE 5
Allowance Spending Regressions

Variable	I Estimates Std. Error	II Estimates Std. Error
Constant	-4802.007 (4571.195)	-1334.858* (424.055)
Margin	2845.194[a] (1571.931)	593.073* (289.371)
Inv	-0.156 (0.106)	-0.066 (0.087)
ROE	4.404 (5.142)	-1.421 (2.392)
Share	2471.164* (766.341)	1289.462 (443.626)
Herf	-0.766[b] (0.453)	-0.153 (0.098)
S_i	6253.155 (6045.509)	1790.114 (3550.556)
Grocery	---	960.258* 329.8612
Software	---	-248.1147 403.3781
Home	---	605.4805 352.4030
Media	---	468.0563 352.2312
Electronics	---	42.3783 359.1579
Pharm	---	360.8744 370.2920
Personal	---	492.6489 374.9083
Clothes	---	432.7828 388.3593

* -- Significant at the 1% level
a -- Significant at the 5% level
b -- Significant at the 10% level

Recall from Table 1 that the three coefficients of interest are those pertaining to a firm's gross margin and market share and industry-level Herfindahl index. The results in Table 5 seem to provide reasonable support for Hypotheses 1 & 2. In the first specification, all three coefficients are of the desired sign and are statistically significant. With the inclusion of industry dummies, the significance of these estimates is somewhat weakened, although all three retain the predicted sign. Gross margin is statistically significant in both specifications.

The coefficient on inventory turnover is negative (although not statistically significant), indicating that firms with higher turnover rates pay smaller promotional allowances. This seems to be reasonable, given our expectations. The results for the firm's perceived profitability (as measured by ROE) are more ambiguous. The coefficient on ROE is not statistically significant in either specification and it actually switches sign when industry dummies are included. I conclude that perceived profitability is not good at explaining promotional spending. The coefficient

corresponding to the grocery industry dummy is positive and statistically significant. This should not be surprising given the prevalence of allowance payments in the grocery industry.

Below, I present the marginal effect estimates from the first-stage binary probit (Table 6):

TABLE 6
Allowance Spending Decision
Probit Marginal Effects (Pay Allowance =1)

	I	II
Variable	**Estimates** **Std. Error**	**Estimates** **Std. Error**
Margin	0.647* (0.262)	0.573[a] (0.255)
Inv	0.000 (0.001)	0.000 (0.000)
ROE	-0.001 (0.002)	-0.001 (0.002)
Share	-0.094 (0.395)	-0.332 (0.488)
Herf	-0.002* (0.000)	-0.002* (0.000)
Grocery	---	0.480[a] (0.205)
Software	---	-0.038 (0.395)
Home	---	0.294 (0.218)
Media	---	0.285 (0.213)
Electronics	---	0.096 (0.308)
Pharm	---	0.330[a] (0.168)
Personal	---	0.396* (0.110)
Clothes	---	0.193 (0.278)

* -- Significant at the 1% level
a -- Significant at the 5% level

Note that the theoretical model makes predictions regarding the magnitude of promotional allowance payments. Hypotheses 1 & 2, therefore, are tested using the second stage results (Table 5). The results presented in Table 6 are not, per se, tests of the theoretical model's predictions. Table 6 does provide some additional insight, however. For example, firms with higher mark-ups are more likely to offer promotional allowances and these allowances are, relatively, large in magnitude. Firms in less concentrated industries are also more likely to offer promotional allowances and offer larger allowances (in dollar amount).

Interestingly, firms in the grocery, pharmaceutical, and personal care industries are more likely to offer promotional allowances. This confirms anecdotal evidence that grocery and drug stores, two main avenues for the sale of pharmaceutical and personal

care products, are most often associated with the use of promotional allowances [Sullivan, 1997]. The remaining coefficient estimates are not statistically significant.

The results in Table 5 appear to provide support for the conclusions drawn from the theoretical model. More specifically, it appears that firms with higher mark-ups and market shares tend to spend more on promotional allowances. Given the interest in studying the possible anti-competitive uses of promotional allowances, this seems to provide at least anecdotal support for those concerns. Additionally, the results indicate that less concentrated industries tend to be associated with higher promotional allowance spending. I believe that this finding relates to the level of competition for in-store promotion and shelf space.

VI. Conclusions

The purpose of this research is to conduct the first analysis of publicly available accounting data on promotional payments. While firms have, traditionally, kept their promotional payment strategies confidential, recent accounting regulation changes allow for a glimpse into the practice. In this paper, I introduce a simple model of retailer-manufacturer bargaining that predicts larger allowance payments from firms with high mark-ups and/or high market share. It also predicts larger allowances in more competitive industries. The results of my two-stage estimation procedure provide support for these conclusions. In addition to these findings, I also conclude that promotional allowances are most prevalent in the grocery, pharmacy, and personal care industries.

It will be interesting to see whether manufacturers and retailers fundamentally change their promotion strategies because of the new accounting rules. Ghitelman [2002] conducts a survey and finds that approximately 62% of supermarkets plan to completely re-evaluate their promotional programs and strategies, while 50% of manufacturers plan a similar re-evaluation. Wellman [2002] finds that there may be other more subtle changes, such as the possible discontinuation of "end-of-quarter blow-out deals," the common promotion strategy wherein manufacturers spend significant amounts of money at the end of fiscal quarters in order to boost sales and revenues. Now that all of that promotion spending will be deducted from net revenues, most industry experts feel it is unlikely that the practice will continue. Unfortunately, there is still a lack of true transparency in SEC financial filings, which makes future research in this area difficult.

References

Anderson, Simon P. and Andre de Palma (2001). "Product Diversity in Asymmetric Oligopoly: Is the Quality of Consumer Goods Too Low?" *Journal of Industrial Economics*, 49 (2): 113 - 135.

Bloom, Paul N., Gregory T. Gundlach, and Joseph P. Cannon (2000). "Slotting Allowances and Fees: Schools of Thought and the Views of Practicing Managers," *Journal of Marketing*, 64 (2): 92 - 108.

Desiraju, Ramarao (2001). "New product introductions, slotting allowances, and retailer discretion," *Journal of Retailing*, 77 (3): 335 - 358.

Ghitelman, David (2002). "Industry Plays New Numbers Game; Accounting Rule Change Forces Manufacturers to Review Allowances," *Supermarket News*, 3 June: 1.

Krishnamurthi, Lakshman and S. P. Raj (1988). "A Model of Brand Choice and Purchase Quantity Price Sensitivities," *Marketing Science*, 7 (1): 1 - 20.

Nagler, Matthew G. (2006). "An Exploratory Analysis of the Determinants of Cooperative Advertising Participation Rates," *Marketing Letters*, 17 (2): 91 - 102.

Sullivan, Mary W. (1997). "Slotting Allowances and the Market for New Products," *Journal of Law and Economics*, 40 (2): 461 - 493.

Wellman, David (2002). "Trade Promos: A Big Chill?" *Frozen Food Age*, 50 (12): 1 - 2.

The Impact of Trademark Counterfeiting On Endogenous Innovation in a Global Economy

Michael W. Nicholson[1]

Assistant Professor of Business Administration, Division of Business and Economics, Transylvania University, E-mail: mnicholson@transy.edu

Abstract:

Optimal policy concerning intellectual property rights, and in particular trademark protection, focuses on the allocation of scarce resources into production or research and development. This paper models intellectual property in two forms—the knowledge of production (or trade secrets) and the reputation for quality in the trademark. Trademark protection that affects counterfeiting achieves a local welfare maximum in the model at an intermediate level between full protection and no protection, slightly lower than that which maximizes innovation. These results are consistent with WTO policies that set minimum international standards but do not necessarily require harmonization.

I. Introduction

Policies that affect firm profits related to intellectual property rights (IPRs) have moved to the center of global trade negotiations over the past twenty years. The Trade-Related Aspects of Intellectual Property (TRIPs) agreement develops international minimum standards for intellectual property protection and makes their enforcement subject to the World Trade Organization dispute settlement mechanism.[2] Trademarks have also been contentious within the recent public debates over globalization, which highlight their role as emblems and instruments of corporate interests.[3] Firms generally appraise the value of a trademark, and the risk of its infringement, with regard to their incentives for entering overseas markets.

This paper introduces trademark counterfeiting to dynamic, general equilibrium models of innovation and growth. The existing literature considers the impact of patent rights and, implicitly, trade secrets, but has not yet discussed trademarks in a setting with innovating firms facing counterfeiting by firms in developing countries. Trademarks differentiate products using signs or marks that identify a firm's reputation, and in some cases the trademark incarnates a large portion of the firm's intellectual property.[4] The loss of this reputation has proven very

[1] I am grateful for the comments and suggestions made by Beata Smarzynska, Christine McDaniel, Amy Glass, James Markusen, Keith Maskus, Jason Moule, Cathy Carey, three anonymous referees, and seminar participants at the Mid-west International Economics Meeting held at Pennsylvania State University.

[2] See Watal (2000).

[3] A critical argument was made forcibly by Klein (2000), with an interesting response from *The Economist* (2001).

[4] See Besen and Raskind (1991).

costly to innovating firms; in a survey taken prior to the adoption of TRIPs, U.S. firms reported that trademarks had great or very great importance in 83% of sales affected by intellectual property, and claimed up to $23.8 billion in lost annual revenues due to its infringement.[5]

The basic justification for IPRs recognizes that innovations are costly to develop, and intellectual property protection enables profit-maximizing firms to recover investments in research and development (R&D).[6] Landes and Posner (1987) associate the various costs of marketing intellectual property using trademarks with the difficulties of achieving excludability in knowledge goods. Landes and Posner contend that the trademark derives its value by lowering search costs and encouraging investments to maintain the quality of trademarked goods. In their model, consumers and firms both gain from the reputation of a trademark, even if its sole purpose simply differentiates goods of identical quality.

Grossman and Shapiro analyze the impact of infringement on investment incentives and property rights in companion models of *deceptive* and *nondeceptive* counterfeiting.[7] Deceptive counterfeiting, developed in Grossman and Shapiro (1988b), refers to the type of trademark infringement analyzed below, in which the violation produces confusion for consumers. Nondeceptive counterfeiting occurs when consumers are likely fully aware of the infringement, such as with $10 Rolexes for sale on Manhattan street corners or $15 Louis Vuitton handbags in Moscow underground markets.

Many authors, notably Helpman (1993), argue that intellectual property rights strengthen the existing monopoly power of advanced countries to the detriment of the developing world. While IPRs may yield dynamic benefits through R&D efforts extended to technological innovation, they produce static losses when potential growth in developing countries is constrained by limited access to existing knowledge. Existing research on patents suggests that IPRs provide necessary incentives for the investment of much industrial R&D; however, similar studies have not yet been conducted for trademarks.[8]

Section 2 of the present paper develops a model in which trademarks signal the quality of vertical innovations, with their value dependent on the level of protection. Innovating firms receive trademarks for quality improvements that become proprietary assets of the firm. Competitors can copy the trademark but cannot imitate the product's quality. Thus, any goods sold under the false mark infringe on the profits of the innovating firm.[9]

Section 3 shows that, under the assumptions of the model, stronger IPRs that reduce trademark infringement in the model will raise the innovation rate in a global economy that experiences much infringement but will lower the innovation rate if protection is already strong.

---

[5] See USITC (1988).

[6] See Maskus (2000) pp 28-33 for a discussion of the impact of IPRs on static and dynamic distortions in the market for knowledge.

[7] See Grossman and Shapiro (1988a,b).

[8] See, for example, Mansfield (1986), Levin, et al. (1987), and Cohen, et al. (2000).

[9] Following convention, I assume perfect protection in the North, where all innovation occurs, with some infringement in the South. IPRs then affect the profits of innovating firms, impact global consumer welfare, and lead to interregional wealth transfers.

Stronger protection of the trademark actually increases the measure of infringed goods if production cannot shift between regions or countries, but may decrease infringement if foreign direct investment (FDI) is possible. The latter result arises because FDI allows fixed Southern resources to flow into production of the latest generation of goods.

Section 4 simulates welfare implications for the model.[10] Global consumer welfare appears to achieve a maximum with an intermediate level of IPRs in the South.[11] The level of trademark protection that maximizes innovation is generally higher than the intermediate level that maximizes welfare, which accounts for the impact of strong protection on market power and subsequent welfare-diminishing price increases. Proper policy on all forms of IPRs must consider the merits of extending monopoly rents, which offers increased incentives for innovation relative to the value of technology diffusion and accompanied lower prices. This paper demonstrates that these considerations should extend to trademark policy and enforcement.

Section 5 concludes.

II. Endogenous trademark infringement

This section develops a model in which innovating firms face the risk of trademark infringement. Southern firms may attempt to sell inferior goods using the established trademark, undermining the value of the symbol and eroding the rents of innovation. IPRs plays a role by reducing the extent to which goods bearing the false trademark can enter the marketplace. [12] The model assumes that firms selling infringed goods ('pirates') can copy the trademark but cannot imitate quality.

The present paper applies an analysis based on vertical preferences, or quality ladders, to trademark enforcement.[13] The quality-ladders framework is a sensible way to model trademark infringement because it can develop the assumption that goods encompass two forms of intellectual property. One form is the *trade secret*, which represents the unique value of any innovation. This production knowledge is the result of innovative R&D efforts, is proprietary to the firm, and cannot be imitated.

When firms innovate a quality improvement, they must signal its value to potential consumers. To do so, they obtain a *trademark* that signals its distinction from previous innovations. This trademark indicates the other form of intellectual property embodied by a

[10] The underlying complexity prevents algebraic derivation, but the simulations demonstrate the expected impacts of changes in levels of trademark control.

[11] The result that asymmetric levels of intellectual property protection would be socially efficient is not uncommon in the literature. See, for example, Scotchmer (2004).

[12] This reflects Article 16.1 of TRIPs that gives the owner of a registered trademark the right to prevent unauthorized use by competitors on similar goods (Watal 2000).

[13] Grossman and Helpman (1991) formalized the model, which has traditionally focused on patents and trade secrets. The fundamental premise of the quality-ladders specification holds that consumers are willing to pay a premium for quality (call it q) for the latest version of a good. They show how the assumption on vertical innovations can lead to a product cycle where production of a good shifts from the North to the South, and is then made obsolete with the introduction of a new generation of quality.

good, its reputation for quality. The premium q consumers are willing to pay incorporates both the value of the good and the reputation for quality.

This paper adopts the convention that after new innovations, the production knowledge incorporated within earlier generations of the product is disseminated throughout the world. The trade secrets are then in the public domain. Firms in the South can produce products up to the penultimate generation, but cannot copy the trade secret of the latest product produced. As introduced below, however, these firms can counterfeit the trademark of the latest good.

2.1 Consumption

Consumption is determined by the following utility function,

$$U_t = \int_0^{\infty} e^{-\rho t} \log u(t) dt, \tag{2.1}$$

where ρ is the discount rate. A continuum of goods exists, indexed by j∈ [0,1]. Firms can innovate new quality levels of each good j to yield a new quality premiums q_m. Thus, log u(t) is determined by

$$\log u(t) = \int_0^1 \log\left[\sum_m q_m(j,t) x_m(j,t)\right] dj, \tag{2.2}$$

where $x_m(j,t)$ indicates consumption of quality m of good j at time t. Quality innovations enter multiplicatively, so that the m^{th} version is valued q more than the $(m-1)^{th}$ version. Consumers purchase the good with the highest quality per price; as shown in the pricing strategy below, this means they purchase only the latest innovation.

Aggregate spending by consumers is given by $E(t) = \int_0^1 \left[\sum_m p_m(j,t) x_m(j,t)\right] dj$. They optimize their lifetime spending at any time t for the instantaneous expenditure E(t). Since all goods enter the utility function symmetrically, instantaneous expenditures are split across all goods evenly.[14]

The intrinsic value of the good as it enters the utility function is the value q_m for the m^{th} innovation. Consumers value the latest good, or m^{th} innovation, at a quality premium q over the penultimate good, the $(m-1)^{th}$ innovation. Since, by assumption, all previous goods are sold at marginal cost w_s, consumers would be willing to pay a maximum qw_s for the latest good.

With full trademark protection, consumers are confident they are buying the latest innovation, and thus are willing to pay the full premium. Most quality-ladder models implicitly assume this signal, but if trademarks can be infringed the signal is imperfect. I assume full

[14] That is, consumers spend E(t)/J for each good, and since J → 1 on the continuum, consumers spend E(t) on every good j at each point in time.

trademark protection in the North and leave a range of potential protection in the South as a policy tool.[15]

2.2 Price decision

Trademark infringement affects both the pricing decision and the market share of innovating firms. Firms engage in price competition to maximize profits. Under the Bertrand assumption, this generally means pricing to capture the full market. All consumers value any quality innovation at q over the last generation of the same good. With a perfect signal, innovating firms can charge q times the production cost of their nearest competitor. Assume all previous innovations are disseminated to the point that production and consumption take place at perfectly competitive prices. Since the nearest competitor faces a marginal cost equal to the Southern wage w_s, innovating firms charge $p^* = qw_s - \varepsilon$ to capture the entire market. As $\varepsilon \to 0$, $p^* = qw_s$. Further assume $qw_s > w_n$ to ensure positive profits for innovators, and for simplicity normalize $w_s = 1$ so that $p^* = q$.

In the presence of trademark infringement, firms are unable to perfectly signal the quality premium. The innovating firm has a monopoly on production of the latest innovation, which they market and sell under the trademark. With infringement, as discussed above, competitors are able to market the penultimate good under the same trademark. The innovating firms thus lose the value of the reputation of their trademark. If the entire market is captured by pirates then innovators face losing all rents, since Southern firms can produce at a lower marginal cost and set prices to take the entire market.[16]

Trademark infringement in the model costs little relative to innovation, but with limited rewards due to the enforcement of existing intellectual property laws. A pirate can copy a symbol or packaging easily but is restricted in its access to the marketplace, since IPRs affect the potential market access for infringed goods. I assume that infringement can only take over a portion of the market, called θ, where $\theta \in [0,1]$.[17] On the aggregate, this is the full market share for a pirate. Innovators whose trademarks have been infringed only lose this portion of their market.

Recall that innovators with a monopoly on the entire market charge $p = q$. Infringing firms pay marginal cost w_s, so they will make a positive profit whenever $p > w_s$. If they charge $p < w_n$, however, they will not sell anything, because no Northern firm would sell below its marginal cost and this low price would signal the low quality of the good. The expected quality premium is $(1-\theta)q$, since q^{m-1} is otherwise sold at $w_s = 1$. Thus, risk-neutral consumers are willing to pay $(1-\theta)q$ for a good sold under the trademark of the latest innovation.[18]

[15] Full Northern protection includes an enforced ban on counterfeit imports. Infringed products manufactured in the South cannot be sold in the North.

[16] There is no infringement of (m-1) goods, since no value accrues to maintaining a trademark on a disseminated product.

[17] Grossman and Shapiro (1988a,b) include a probability of counterfeit products being confiscated. The present assumption could reflect that certain portions of the physical markets for goods are avoided by pirates due to the higher risk of confiscation in those areas.

[18] Risk-averse consumers would be willing to pay less, by Jensen's inequality. Alternatively, the θ parameter could capture risk in the model. This assumption does not affect the results.

Suppose the innovator sets the original price, and all pirates must follow its lead. Since consumers are not willing to pay any price above the maximum (1-θ)q, no separating equilibrium is possible, and thus the innovators have the incentive to charge $p^* = (1-\theta)q$. At this price, any attempt by a pirate to steal the market with a lower price would signal the poor quality of their good. Thus, p^* holds as an equilibrium price.[19]

2.3 Profit equations

Successful innovators obtain a monopoly on the production of the latest innovation as well as its trademark. Three firm types exist: firms with an uninfringed mark (indexed by N); firms whose marks have been infringed (indexed by NT); and pirates (indexed by T). For uninfringed firms, the trademark works as a perfect signal, so consumers with expenditure level E pay the maximum price for the good $p^N=qw_s$. These firms sell quantity E/q, and pay marginal cost w_n. Let $w_s=1$ and define the relative wage $w = w_n/w_s$, which leads to the following profit equation:

$$\pi^N = x_j(p_j - w_n) = \frac{E}{qw_s}(qw_s - w_n) = \frac{E}{q}(q - w). \tag{2.3}$$

After successful infringement, pirates obtain a maximum market share θ in the South. Potential infringement occurs at intensity τ, which is the rate chosen by the pirate that resources are expended to the task. The probability that a trademark will be infringed is τ after which infringed goods make up a portion θ of the market. The value of τ arises endogenously in the steady-state equilibrium.

I assume that once a trademark has been infringed Southern consumers on the aggregate are instantaneously aware of this (although not which particular goods are infringed) and adjust their behavior accordingly.[20] Define s to be the share of global income controlled by the Southern consumers. Pirates charge $p^T = (1-\theta)q$, sell quantity $\theta\frac{sE}{(1-\theta)q}$, and pay marginal cost $w_s=1$, to earn profits

$$\pi^T = \frac{\theta}{1-\theta}\frac{sE}{q}((1-\theta)q - 1). \tag{2.4}$$

[19] As an example, consider Sony radios being sold in Bangkok, with a marginal cost w_S for the latest model \$10 and a quality increment q of 3.5. Consumers are thus willing to pay \$35 for this year's model. Now suppose the Sony trademark is infringed to share θ=0.286. As soon as the first inferior product is sold under the brand Sony, consumers on the aggregate lower their willingness to pay to $(1-\theta)qw_S$, or \$25. Sony not only loses market share but also a portion of its price mark-up.

[20] This assumption is close to Allen (1984)'s conjecture that buyer's evaluations of purchases in period t are known to all consumers prior to making purchases in period t+1.

Northern firms whose products have been infringed (indexed by NT) also charge $p^{NT}=(1-\theta)q$ in the South and sell quantity $(1-\theta)\frac{sE}{(1-\theta)q}$, charge q and sell quantity $(1-s)\frac{E}{q}$ in the North, and pay marginal cost w to earn profits

$$\pi^{NT} = \frac{E}{q}(q-w) - s\theta E . \tag{2.5}$$

2.4 Research and development

Northern firms invest resources at intensity ι to innovate quality improvements, with the labor cost of innovation I. The R&D cost behind an innovation is wIdt, since Northern labor is engaged in the activity, with the expected gain $\iota v^N dt$, where v^N gives the present value of an innovation. Southern firms invest resources at intensity τ with labor cost T to infringe on trademarks of existing innovations. Successful pirates gain $\tau v^T dt$ at cost Tdt.

I assume free entry in innovation and infringement, so that the expected gain cannot be greater than the cost for either R&D equation. The equations hold with equality for positive rates of R&D activity, which leads to the following expressions:

$$v^N = wI \tag{2.6}$$

$$v^T = T . \tag{2.7}$$

2.5 No-arbitrage

Under the rational-expectations stock market valuation developed by Grossman and Helpman (1991), individuals invest in firms until they reach the same expected value as a riskless bond earning interest r times the value of the firm. Northern firms earn the profits $\pi^N dt$, with capital gain $\dot{v}^N dt$. With infringement, Northern firms lose the expected value $\tau(v^N - v^{NT})$. They also face the risk of capital loss $\iota v^N dt$, in which other firms innovate over their quality. This yields the following no-arbitrage condition:

$$\dot{v}^N + \pi^N - \iota v^N - \tau(v^N - v^{NT}) = rv^N . \tag{2.8}$$

Dividing through by v^N yields

$$\frac{\pi^N}{v^N} + \frac{\dot{v}^N}{v^N} - \tau\frac{v^N - v^{NT}}{v^N} - \iota = r . \tag{2.9}$$

Along the balanced growth path that $\frac{\dot{v}^N}{v^N} = 0$ and ρ = r, (2.9) can be rearranged to be

$$v^N = \frac{\pi^N + \tau v^{NT}}{\rho + \iota + \tau}. \tag{2.10}$$

Similar analysis yields the following expressions for v^{NT} and v^T:

$$v^{NT} = \frac{\pi^{NT}}{\rho + \iota} \tag{2.11}$$

$$v^T = \frac{\pi^T}{\rho + \iota}. \tag{2.12}$$

2.6 Resource constraints

Production and R&D efforts are constrained by the scarce resources available to both regions in the model. Northern labor is used for innovation by firms engaging in R&D, and in production by N and NT firms. Define n_j as the measure of firm types, on a spectrum of 0 to 1. The Northern firms with recent innovations, of measure n_N, produce E/q goods for a total labor use of $n_N E/q$. Northern firms with an infringed trademark, of measure n_T, also produce E/q goods for a total labor use of $n_T E/q$.[21] Firms that are engaging in research to achieve new innovations expend ιI units of labor on the full continuum of goods. These lead to the following expression of the Northern resource constraint:

$$L_N = \iota I + n_N \frac{E}{q} + n_T \frac{E}{q}. \tag{2.13}$$

In the South, firms engage in infringement at labor cost $n_N \tau T$ and sell quantity $\frac{sE}{(1-\theta)q}$ to proportion $n_T \theta$ of the market, for a resource constraint

$$L_S = \tau n_N T + n_T \frac{\theta}{1-\theta} \frac{sE}{q}. \tag{2.14}$$

2.7 Constant measures

In the steady-state, the measures of every firm type must remain constant. That is, the number of firms that become pirates must be equal to the number of firms who stop being

[21] Note that $n_{NT} = n_T$, since for every infringed firm there exists an equivalent infringing firm. These infringed firms are actually selling $\frac{E}{(1-\theta)q}$ goods to (1-θ) of the Southern market (or s(1-θ) of the total market) and E/q goods to the Northern market (or (1-s) of the total market) for total production E/q.

pirates. Since innovation occurs on all types of firms, at any given time the number of firms becoming innovating firms is $\iota(n_N + n_T) = \iota$. Firms are no longer innovating firms after infringement or after another innovation, thus the measure of firms leaving n_N is $(\iota+\tau)n_N$. Firms become pirates through infringement on the measure n_N and leave through innovation on the measure n_T.

In the model, I wish to consider the effects of trademark protection on the endogenous variables ι and n_T, which focuses the analysis to the primary concerns of intellectual property protection: the innovation rate and the extent of infringement. The following equations show the relationships between (ι, n_T) and (τ, n_N) that help reduce the structural equations to the elements in question:

$$n_N = 1 - n_T \tag{2.15}$$

$$\tau = \iota \frac{n_T}{1 - n_T}. \tag{2.16}$$

III. Analysis

3.1 Intuitive description of reduced-form equations

To conserve on mathematical complexity, Appendix A.1 develops reduced-form equations for equations of the balanced growth path developed in Section 2. This section provides an intuitive description of the effects of trademark protection based on these reduced-form equations by graphing the relationship of the innovation rate, ι, and the measure of pirated firms, n_T, within the resource constraints. By solving for E/q from (A.2) and plugging into the resource constraints (A.3) and (A.4), two equations for ι and n_T are given to be:

$$L_N = \iota I + \frac{(\rho + \iota)T}{s\theta(q - \frac{1}{1-\theta})} \tag{3.1}$$

$$L_S = m_T T + n_T \frac{(\rho + \iota)T}{(1-\theta)q - 1}. \tag{3.2}$$

Define the results for the Northern resource constraint "LN" and the Southern resource constraint "LS". Fully differentiating (3.1) and (3.2) gives an expression for $dn_T/d\iota$ for both resource constraints.

$$\text{LN: } \frac{dn_T}{d\iota} = 0 \tag{3.3}$$

and

$$\text{LS: } \frac{dn_T}{d\iota} = -\frac{n_T(1-\theta)q}{\iota(1-\theta)q+\rho} < 0\,. \tag{3.4}$$

The relationship between the innovation rate and the measure of infringed goods can be identified in (ι, n_T) space. The LN line is vertical at a point ι*, according to (3.2), and the LS line is upwardly concave since the second derivative of (3.3) is positive. The measure of infringing firms does not affect the resources in the North, so the innovation rate ι* is determined independently of n_T.[22]

The value of ι* can be solved from (3.1) to be

$$\iota^* = \frac{sL_N\theta(q-\frac{1}{1-\theta})-\rho T}{sI\theta(q-\frac{1}{1-\theta})+T}\,. \tag{3.5}$$

Figure 1 provides an illustration of the effects of trademark protection on the endogenous variables ι and n_T. Consider first the LN line. Since n_T is not present in the equation, ι is the only variable of concern. From (3.5), the derivative of ι* with respect to θ is

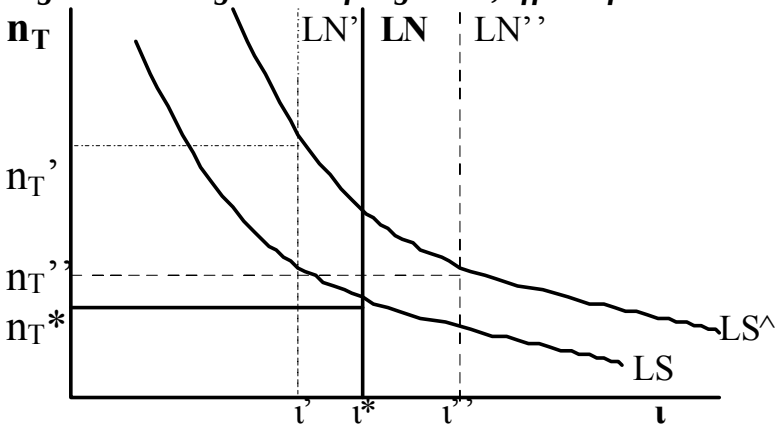
$$\frac{d\iota^*}{d\theta} = \Psi\frac{s\rho L^N IT}{[I\theta(q-\frac{1}{1-\theta})+T]^2}\,. \tag{3.6}$$

The value of this derivative depends on the value of the term Ψ, a combination of q and θ that also determines the signs of dE/dθ and dι/dθ. If ψ is positive then the innovation rate rises with IPRs and the LN line shifts right, but if it is negative then the innovation rate decreases and the LN line shifts left.

[22] This holds because Northern resources in innovation draw from production by the total output of n_N and n_{NT} firms, which is not affected by n_T. The n_N firms produce $\frac{E^N+E^S}{q}n_N$ and the n_{NT} firms produce $\frac{E^N}{q}n_T+\frac{E^S}{q(1-\theta)}(1-\theta)n_{NT}$, which gives total constant Northern production to be E/q. Real quantity output per firm in the North is the same for infringed and uninfringed firms since the price discount (1-θ) exactly offsets the limited market share. The measure n_{NT} (and its equivalent n_T) cancels out in the Northern resource constraint.

Similarly, IPRs affect the LS line through (3.2). As θ decreases, the right-hand side of (3.2) decreases, so that higher values of both ι and n_T are required for the equation to hold. This implies a shift right in the LS line. It can be shown that $dn_T/d\theta < 0$, so that the measure of infringing firms increases regardless of effect on ι*. This is reflected in Figure 1, where n_T^* is lower than either n_T', when LN shifts to the left, or n_T'', when LN shifts to the right.

Figure 1: Endogenous infringement, effect of IPRs



3.2 Summary of Results

Three major results concerning the general equilibrium impact of IPRs arise from the model to this point.

Proposition 1: *For large values of θ, implying initially weak protection of intellectual property, the innovation rate ι increases with stronger IPRs. For small values of θ, ι declines in IPRs. The innovation rate achieves a maximum at the point where* $\Psi = 0$, *or where* $\theta = 1 - q^{-1/2}$.

The level of trademark protection that maximizes innovation occurs at an intermediate level. Intuitively, this reflects the tradeoffs between resources engaged in R&D and production. The impact of changes in IPRs may increase or decrease the innovation rate depending on characteristics of the economy. For large values of θ relative to q, then $d\iota/d\theta < 0$, and an increase in IPRs leads to an increase in the innovation rate. For small values of θ relative to q, however, an increase in IPRs causes the innovation rate to decline. This implies that non-innovating countries may be discouraging innovation with either very strong or very weak protection of trademarks.

These results must be interpreted, however, in the context of the welfare implications discussed in Section 4. Utility does not necessarily achieve a maximum when the innovation rate is greatest, since the level of production reaches a minimum at that point. Scarce resources must be split between two utility-generating activities—production and R&D—which leads to a clear trade-off in the optimal value of IPRs.

Proposition 2*: The measure of infringed goods increases in IPRs.*

With less infringement fewer resources are devoted to the production of infringed goods, which raises the rate of infringement and the measure of infringed goods. With stronger trademark protection, the measure of goods with infringed trademarks increases. A greater percentage of goods now face infringement, despite stronger policies denying infringement. This result arises because trademarks lower the actual market share of counterfeit goods.

Since stronger IPRs lead to a market share decline for each infringed good, Southern resources shift to a broader spectrum of goods. Thus, production per good declines, but the overall measure increases. The assumption driving this result is that fixed Southern resources are engaged either in infringement or production of infringed goods. If production of innovated goods can be shifted to the South by FDI, then this result may be overturned for some values of the parameter space.

Proposition 3: *The Northern relative wage increases as IPRs strengthens for values of θ near zero.*

Although the measure of infringed goods grows with IPRs, according to Proposition 2, the actual resources used in infringement are very small for θ near zero. The increase in n_T is negligible in this range, so overall infringement decreases. With less infringement, labor demand shifts to the North, so w increases.

IV. Welfare effects

Social welfare in the model derives entirely from consumer utility, since free-entry in research and development implies no dynamic firm rents. The quality-ladders specification (2.2) provides three sources of utility: 1) amount of goods produced, represented by aggregate expenditures E; 2) the entire vector q^m embedded in a good; and 3) prices. This section discusses how these elements interact with IPRs, and the results suggest that global welfare achieves a maximum at intermediate levels of protection.

The set of equations (A.5) – (A.8) are sufficiently complicated in θ that an algebraic derivation cannot be obtained without substantial simplifying assumptions. These assumptions would render the analysis of trademark protection impractical. Thus, this section presents numerical simulations that demonstrate the expected impact of changes in policy and enforcement.

4.1 Infringement and welfare

Given the instantaneous utility function in (2.2), consumers spend value E(t) on every good available and, due to the price competition, purchase only one variety of each product. [23]

[23] Since the elasticity of substitution is 1, consumers split their nominal spending evenly across all products, so that E(j,t) = E(t)/J. On the continuum that J→1, then, E(j,t) = E(t).

Call this quality level $\tilde{m}(j,t)$. The amount purchased is given by $x_{\tilde{m}}(j,t) = \dfrac{E(t)}{p_{\tilde{m}}(j,t)}$, which yields the following expression for instantaneous utility:

$$\log u(t) = \int_0^1 \log\left[\sum_m q_{\tilde{m}}(j,t)\frac{E(t)}{p_{\tilde{m}}(j,t)}\right]dj . \tag{4.1}$$

Since only the highest quality is sold, $\sum_m q_{\tilde{m}}(j,t) = q^{\tilde{m}(j,t)}$. E(t) is independent of j, giving

$$\log u(t) = \log E(t) + \int_0^1 \left[\tilde{m}(j,t)\log q - \log p_{\tilde{m}}(j,t)\right]dj . \tag{4.2}$$

Let $\bar{m}$ and $\bar{p}$ denote the average quality-levels and prices, respectively, across all products. Integrating across these averages in (4.3) gives

$$\log u(t) = \log E(t) + \bar{m}\log q - \log \bar{p} . \tag{4.3}$$

To find $\bar{m}$, consider the expression $\log q^{\tilde{m}(j,t)} = \tilde{m}(j)\log q$. The uninfringed products sold, of measure n_N, have quality q^m. The infringed products sold, of measure n_T, have quality q^{m-1}. This yields

$$\log q^{\tilde{m}(j,t)} = [n_N\,\tilde{m} + n_T(\tilde{m}-1)]\log q = (\tilde{m} - n_T)\log q \tag{4.4}$$

By the law of large numbers, $\tilde{m} = \iota$, and with the integration of (4.4), $\bar{m} = \iota - n_T$.[24]

The prices of each type of good determine $\bar{p}$. Non-infringed goods sell at price q, infringed goods at price (1-θ)q, giving

$$\bar{p} = \frac{1}{J}\sum_j p_j = (1-n_T)q + n_T(1-\theta)q = q(1-\theta n_T) . \tag{4.5}$$

Plugging $\bar{m}$ and $\bar{p}$ into (4.4) yields

[24] See Grossman and Helpman (1991), page 97, and Glass and Saggi (2002).

$$\log u(t) = \log E(t) + (\iota - n_T)\log q - \log q(1 - \theta n_T)\,. \tag{4.6}$$

This expression rises with E and ι but generally falls with n_T. As aggregate expenditure (production) and innovation rise, or the measure of infringement falls, welfare increases, all intuitive effects. The expression (4.6) can be solved by plugging in (3.6) – (3.8). Recall that n_T is always decreasing in IPRs, but the effects of IPRs on E and ι depend on the sign of ψ.[25] The interactions of these terms yield an inverted-U shape for welfare. As can be shown, θ^* that maximize innovation is less than the welfare-maximizing θ, which means that welfare achieves a maximum when IPRs are *weaker* than that which maximizes innovation.[26]

Thus, welfare achieves a maximum in this model at an intermediate level of IPRs in the South lower than that which maximizes innovation. Intuitively, this result recognizes that the intermediate level of protection represents the trade-off between production and R&D in the model. Fixed resources must be allocated to welfare-enhancing activities, and maximum innovation occurs at the same level of the policy parameter as minimal production. The welfare-optimal policy relaxes Southern trademark protection sufficiently to shift resources into production.

An interesting implication of this result is its similarity to current levels of IPRs. In the present model, trademarks are perfectly protected in the North, but welfare achieves a maximum when Southern IPRs are relaxed enough to allow some infringement. This reflects real-world levels of implementation.

V. Conclusion

Trademark counterfeiting continues to occur on a very regular basis throughout the world. On a recent trip to the former Soviet Union, the author casually discovered widespread – and obvious – piracy of global trademarks. For example, a local fast-food franchise in Yerevan, Armenia advertised a "popcorn chicken" replete with full replication of *Kentucky Fried Chicken*'s trademarked Colonel Sanders; this image was removed from the advertising prior to the arrival of an officially-licensed *KFC* franchise. Official discussions with multiple corporate investors for various Western companies in Yerevan expressed a clear reluctance to seek business opportunities within the former Soviet Union specifically due to such infringement.

International differences in intellectual property rights have become a focus of intense multilateral negotiations in forums such as the World Trade Organization. The complexity of the subject demands sophisticated research regarding optimal policy recommendations. This paper demonstrates that in a dynamic, general equilibrium model of the global effects of intellectual property protection that trademark counterfeiting may yield some positive impact on welfare. The leading concerns for policy on IPRs, according to this model, include the allocation of scarce resources into production or R&D.

---

[25] From above, $\Psi = \left(q - \frac{1}{(1-\theta)^2} \right)$

[26] Full derivations of the mathematical proofs can be found in Nicholson (2006), an earlier version of the current paper.

Section 2 extends the traditional consideration for IPRs beyond the general equilibrium effects of increasing imitation costs, recognizing that consumers are motivated by the value of the firm's reputation. Section 3 shows that the maximum innovation rate occurs when the Southern (non-innovating) region sets IPRs that allow an intermediate level of trademark infringement. Welfare, however, achieves a maximum at a level of IPRs slightly weaker than the one that maximizes innovation. As Section 4 demonstrates, at this point resources in the Northern (innovating) region are efficiently allocated between production and R&D.

The results of this model are consistent with policies adopted by the TRIPs agreement, which does not require policy harmonization (as is popularly believed) but sets minimum international standards. For example, according to Article 14.3 of the agreement, individual countries may make registration of a trademark depend on its use by the applicant, but are limited by the scope of those requirements. The theory and simulations presented in the paper have not been quantified, however, and the suggested welfare-optimal level of trademark protection may be the level currently adopted by TRIPs. As with most theoretical discussions regarding intellectual property rights, the clear next step is to seek rigorous empirical estimation of the above propositions.

References

Allen, Franklin (1984). "Reputation and Product Quality," *Rand Journal of Economics*, 15 (3): 311 - 327.

Besen, Stanley and Leo Raskind (1991). "An Introduction to the Law and Economics of Intellectual Property," *Journal of Economic Perspectives*, 5 (1): 3 - 27.

Cohen, Wesley M., Richard R. Nelson, and John P. Walsh (2000). "Protecting Their Intellectual Assets: Appropriability Conditions and Why U.S. Manufacturing Firms Patent (Or Not)," NBER Working Paper No. 7552.

Glass, Amy and Kamal Saggi (2002). "Intellectual Property Rights and Foreign Direct Investment," *Journal of International Economics*, 56: 337 - 360.

Grossman, Gene and Elhanan Helpman (1991). *Innovation and Growth in the Global Economy*, Cambridge, Massachusetts: The MIT Press.

Grossman, Gene and Carl Shapiro (1988a). "Foreign Counterfeiting of Status Goods," *The Quarterly Journal of Economics*, 103 (1): 79 - 100.

Grossman, Gene and Carl Shapiro (1988b). "Counterfeit-Product Trade," *The American Economic Review*, 78 (1): 59 - 75.

Helpman, Elhanan (1993). "Innovation, Imitation, and Intellectual Property Rights," *Econometrica*, 61 (6): 1247 - 1280.

Klein, Naomi (2000). *No Logo*, United States: Picador USA.

Landes, William and Richard Posner (1987). "Trademark Law: an Economic Perspective," *The Journal of Law and Economics*, 30: 265 - 309.

Levin, Richard C., Alvin Klevorick, Richard R. Nelson, and Sidney G. Winter (1987). "Appropriating the Returns from Industrial Research and Development," *Brookings Papers on Economic Activity*, SP ISS: 783 - 820.

Mansfield, Edwin (1986). "Patents and Innovation: An Empirical Study," *Management Science*, 32 (2): 173 - 181.

Maskus, Keith (2000). *Intellectual Property Rights in the Global Economy*, Washington, D.C.: Institute for International Economics.

Nicholson, Michael (2006). "Trademark Infringement and Endogenous Innovation," University of Colorado Working Paper 00-12.

Scotchmer, Suzanne (2004). "The Political Economy of Intellectual Property Treaties," *Journal of Law, Economics, and Organizations*, 20: 415 - 437.

"The Case for Brands," *The Economist*, 6 September 2001.

U.S. International Trade Commission (1988). *Foreign Protection of Intellectual Property Rights and the Effect on U.S. Industry and Trade.* USITC Publication 2065.

Watal, Jayashree (2000). *Intellectual Property Rights in the World Trade Organization: The Way Forward for Developing Countries.* New Delhi: Oxford University Press, India and London: Kluwer Law International.

Appendix A

A.1 Reduced-form equations

The equations from Section 2 can be combined to provide insight into the economy. Combine the profit equations (2.3) and (2.5) and the value equations (2.10) and (2.11) with the R&D equation (2.6) and the constant measure (2.16) to get

$$\frac{E}{q}\left[q - w - \frac{sq\theta\iota n_T}{\rho(1-n_T)+\iota}\right] = (\rho+\iota)wI\,. \tag{A.1}$$

In similar fashion, combine the profit equation (2.4) and the value equation (2.12) with the R&D equation (2.7) to get

$$\theta\frac{sE}{q}\left[q - \frac{1}{1-\theta}\right] = (\rho+\iota)T\,. \tag{A.2}$$

Equation (A.1) indicates the condition of zero economic profits for innovating firms. The left-hand side shows the profits for successful innovations, and the right-hand side shows the cost of innovation weighted by the discount rate and the risk of capital loss. Equation (A.2) indicates a similar relationship for infringement.

Plug (2.15) and (2.16) into (2.13) and (2.14) to get resource constraints in terms of ι and n_T:

$$L_N = \iota I + \frac{E}{q} \tag{A.3}$$

and

$$L_S = \iota n_T T + n_T\frac{\theta}{1-\theta}\frac{sE}{q}\,. \tag{A.4}$$

Equations (A.1) – (A.4) provide four relationships for the four endogenous variables {E, w, ι, n_T}. The effects of trademark protection on these variables can be derived from Cramer's Rule.[27] Equation (A.1) solves the relative wage in terms of the other three:

$$w = \frac{E\left(1 - \frac{s\theta\iota n_T}{\rho(1-n_T)+\iota}\right)}{\frac{E}{q} + I(\rho+\iota)}\,. \tag{A.5}$$

[27] Full derivation provided in Nicholson (2006).

Similarly, (A.2) - (A.4) solve for the other variables. For simplicity, define $Q \equiv 1 - \frac{1}{q(1-\theta)}$ to get

$$E = \frac{q(\rho + \iota)T}{s\theta Q} \tag{A.6}$$

$$\iota = \frac{L_N - E/q}{I} \tag{A.7}$$

$$n_T = \frac{L_S}{\iota T + \frac{\theta}{1-\theta}\frac{sE}{q}}. \tag{A.8}$$

These equations can be solved to show that $dw/d\theta < 0$ for small θ, so the relative wage increases as IPRs approach perfect protection.[28] It can also be shown that the measure of pirating firms is declining in trademark protection (as would be expected), but that $dE/d\theta$ and $d\iota/d\theta$ have opposite signs which depend on the level of infringement θ and the size of quality improvements q, specifically the sign of $\left(q - \frac{1}{(1-\theta)^2}\right)$, henceforth defined to be Ψ. If $\Psi < 0$, then $dE/d\theta > 0$ and $d\iota/d\theta < 0$, and stronger trademark protection leads to an increase in the innovation rate. The necessary condition for pirates to enjoy positive profits is $q > \frac{1}{1-\theta}$, but there is no clear relationship between q and $1/(1-\theta)^2$. For low levels of infringement, a weakening of IPRs that raises the infringement rate will increase the innovation rate even with quality levels near unity. In this region, weaker IPRs increase innovation. If $\psi>0$, with high infringement levels and sufficiently low quality improvements, stronger IPRs increase innovation.

This implies that the innovation rate achieves a maximum along the line where $\psi=0$, or where policy on intellectual property sets $\theta = 1 - q^{-1/2}$. At this point, infringement levels and quality improvements combine to provide optimal returns to product innovation. Note that q is a fixed parameter, relating the value of any innovation, and θ is a policy parameter dependent on the strength of IPRs. Thus, for every given q, there exists a θ such that the innovation rate achieves a maximum. Production, however, achieves a minimum along the same set of points, and as discussed in Section 4 below, this is not necessarily the point where welfare achieves a maximum.

[28] Recall that an increase in the strength of IPRs indicates a decrease in θ. Although explicit conclusions cannot be drawn for θ not close to zero due to the complex relationships of the variable, simulations in section 4 provide general conclusions about solutions to the model.