Journal of **Economics and Finance Education**

Volume 8 **Winter, 2009** **Number 2**

Attraction and Retention of Faculty in a Non-Tenure Granting Environment

Abstract

This paper examines the task of attracting and retaining faculty in a non-tenure-granting environment. Our goal is to provide a framework for the future discussion of this issue and to develop a set of utility functions for the primary parties involved, namely, institutions of higher education and individual faculty members. Various strategies for attracting and retaining faculty members in a non-tenure-granting environment are evaluated in terms of their long-run implications.

Introduction

Tenure in institutions of higher education is under going increasing scrutiny. While institutions show little inclination to try to revoke the tenure status of those who already hold it, the conditions of contract for incoming faculty are clearly changing. While the overall number of full-time faculty positions has increased, t he granting of tenure has not increased. Data provided by the American Association of University Professors (AAUP) indicate that, in 2006, approximately 53.5% of all full-time faculty held tenure, approximately the same percentage that held it in 1975. During the same period, however, the number of full-time faculty on tenure track fell from 29% to 23% and the percentage of full-time professors in a non-tenure granting or contract systems increased from 19% to 23.4%.

These changes in hiring practices reflect prevailing market conditions. In the labor market, a relative surplus of Ph.D.'s in many fields has allowed institutions to reduce the value of the employment package offered to many incoming faculty. At the same time, institutions of higher education have been called to be more responsive to their stakeholders. Students, the tax-paying public, funding agencies, and the organizations that hire students have all placed new pressures on higher education: two major issues are the costs of higher education and the flexibility. The issue of flexibility concerns both offerings and staff. The flexibility of offering's issue is a major driver behind the move towards distance-learning and flextime offerings. Flexibility of staff issues are driven by recognition of the long-term commitment which tenure entails and recent attempts to change or update traditional curricula where tenured faculty members are resistant to changes in a curriculum to which they are vested. During the same period

of time, the demographic base of students has shifted away from the "traditional" student towards "non-traditional" students and growth rates in the price of tuition (relative to other commodities) have continued to climb.

All of these factors have contributed to increased scrutiny of the role and effects of tenure among college and university faculties. Higher education is clearly facing an era where external pressures to control costs are real (often from state legislative bodies), and where faculty salaries represent a major cost component.

The results of these phenomena are felt at most institutions. While a plethora of descriptive statistics support this contention, the authors have yet to find a systematic economic analysis of the factors that affect the offer of tenure in contracts. In order to explore these changes more systematically, this paper will focus on the utility functions for the individual faculty member and for the institution of higher education. A human capital approach is employed. Various strategies for attracting and retaining faculty members are evaluated in terms of short run and long run effects for faculty and for institutions.

Attraction and Retention of Faculty - Faculty Perspective

The factors that initially attract faculty are the same that affect retention (the later is defined as the decision to remain in or leave an institution), We posit the following utility function for individual faculty:

$$Utility = f(W, RE, TE, SE, IC, P)$$

Where: W = Wage
RE = Research Expectations
TL= Teaching Expectations

SE = Service Expectations
IC = Institutional Commitment
P = a vector of factors impacting personal utility

Wage

A combination of general labor market conditions and the reputation or vitae of the individual are the primary determinants for the wage of a faculty member. Our framework focuses on the differences between tenured and non-tenured environments. Therefore, we want to more clearly delineate the components of the offered wage. We posit the wage offer to be the following:

$$Wage = f(W_{rf}, CD_1, CD_2, CD_{3,\wedge}, CD_n)$$

Where: W_{rf} = Risk-free Wage
CD_n = Compensating Differentials

$$1 \rightarrow n$$

The risk free wage is the wage that prevails for a given set of skills ignoring compensating differentials found in job market. Compensating differentials might include: risk of injury, risk of death, fringe benefits, job status, job security, job location, extent of control over one's work environment, potential for growth in wage, among others. These compensating differentials affect the wage offer. Compensating differentials which the employee view's as favorable will decrease the wage offer and compensating differentials which the employee view's as unfavorable will increase the wage offer.

While all of these differentials have the potential to affect wage some of them are specific to the job, while others are specific to a particular position. For instance, it is widely accepted that the flexibility embodied in academic posts represents a positive compensating differential and this accrues to the job generally. For our purposes here we wish to ignore compensating differentials which accrue to the job generally and to focus on the compensating differential for job security as it relates to tenure practices.

When a given job offers protection from unemployment (e.g., it offers tenure) this reduces the risk of an interruption of an employees expected earnings stream. In a typical scenario, the faculty member comes into the institution under a probationary (non-tenured) condition. This set of conditions assigns the acceptance of risk to the faculty member and a compensating differential for that risk is awarded in the market. Thus, for the pre-tenure employment contract:

$$\partial W / \partial R \geq 0$$

where R represents the job security risk

Over the life of the employment contract, this risk premium will fall. Once tenure is awarded the risk premium should approach zero. In a post-tenure contract where job security is assured, the value of the risk premium becomes negative.

$$\partial W / \partial R \geq 0$$

To summarize, we expect the risk premium to be positive when a faculty member is hired into a position but that it will fall over the length of the employment contract and will become negative in a post-tenure position.

This framework can help to explain observed salary compression across ranks and the economic value of mobility in the market. There are clearly cases where some institutional constraint on wage increases contributes to the decision by faculty to move into the job market in order to recalibrate their wage against that market.

Other Factors and Implications for Faculty Members and Institutions

Given the predominance of tenure in the academic job market and the nature of the risk premium outlined above, we expect initial wage offers among both tenured and non-tenured positions to be posted to the market-clearing wage. During the period of the employment contract, the compensating differential for risk is maintained, though at a decreasing level over time, in both environments[1].

Upon the granting of tenure we maintain that the job security risk premium disappears. While we may observe a relatively substantial increase in salary at tenure, that increase is related to promotion and is reduced by a job security compensating differential that has become negative.

This differs substantially in a non-tenured environment where the initial compensating differential for risk must be either maintained throughout the length of the employment contract or reinstated upon contract renewal if an institution wishes to retain mobile employees.

The withdrawal of the offer of tenure also serves to reduce the gap between academic and non-academic jobs for individual faculty. The authors expect salaries in a non-tenured environment to more closely approximate wages outside of academe over the long term. Faculty salaries are already affected by opportunity costs and if non-

[1] The initial compensating differential for job security risk must represent the sum of the net present values of the expected year-by-year compensating differential, across the length of the pre-tenure contract.

tenure environments grow relative to tenured environments, this problem is likely to be exacerbated. This effect will differ across disciplines and be most pronounced where opportunity costs are highest.

Institutional Commitment will play a critical role in the behavior of faculty. If faculty view renewal decisions as closely linked to their fulfillment of the mission of the institution and they are comfortable with the outcomes of renewal and non-renewal decisions, this factor will tend to lose importance. However, if faculty view these decisions as arbitrary, unfounded, or unfair, this factor will gain importance. Under the later set of conditions, mobility will become a key goal for faculty. If connections to the institution are viewed as tenuous, individual faculty will pursue a career path that maximizes their mobility within the profession. Faculty will maintain a level of research, teaching, and service that will fit their expectations of the requirements needed to gain a new position.

For the institution, decisions to renew become critical for two major reasons. First, if faculty view renewal decisions negatively, they will pursue an agenda that maximizes their individual mobility within the profession. Faculty members who are pursuing mobility are more likely to be engaged in research because mobility within the profession is influenced primarily by research productivity. If research is the primary mission of the institution, this is not an immediate problem -- faculty continue to pursue a self-interest agenda which matches that of the university. If teaching is the primary mission of the institution, the agenda's collide. This is important because many of the institutions that offer non-tenure contracts include teaching as a major component of their

stated mission. For these institutions, faculty fit will diminish if faculty pursue a mobility agenda.

Second, over the longer term, reputation will affect the institution's ability to attract candidates, particularly at the associate and full professor levels, where information about institutions is more widely known and shared. Reputation effects that are perceived negatively in the market would reduce the pool and/or quality of faculty candidates. The pool is likely to differ by rank because the level and quality of information increases for faulty who have stronger network information -- and we expect information about the job market to increase with experience in it.

Faculty willing to accept the increased risk of a non-tenured slot may differ from the general labor pool. The utility functions of those with tenure and those without tenure are likely to differ substantially. Hence, the applicant pool of faculty might be divided into individuals with tenure and individuals without tenure. For faculty members leaving a tenured slot, it seems plausible that non-monetary factors predominate. These individuals are seeking an increase in psychic income relative to real income. Factors such as: preference for location, the ability to change personal or work environment, or the opportunity to engage in a more entrepreneurial venture are likely to play an expanded role. For instance, if an institution is new when it hires faculty on non-tenure lines (e.g., Evergreen State College in 1971 and Florida Gulf Coast University in 1997), there is little information provided in the market concerning working conditions or institutional commitment, so risk is higher than usual. One factor that might attract faulty to these institutions is a belief that a new institution offers more opportunities for entrepreneurial endeavors within the academy. For faculty interested in building a

particular program or institution, the opportunity to escape traditional constrictions that are inherent to long-established programs may be a major draw.

Non-tenured institutions may also attract individual faculty who simply reject the excesses of the tenure system that he/she has observed over his/her career. These individuals see the connections between academic freedom and tenure as weak and are willing to trade-off job security for what they perceive to be an enhanced work environment.

Another category of tenured faculty attracted to a non-tenured slot would be experienced faculty, fully-vested in a retirement plan who are seeking a change for some reason. Again, these individuals are likely to be more strongly influenced by non-monetary factors and seeking an increase in psychic income relative to real income.

The second major group is faculty without tenure. The first category within this group is composed of newly-minted Ph.D.'s. These individual tend to have less market information and their willingness to accept a non-tenured slot may be connected to a lack of offers of tenured positions[2]. A second category within this group is composed of faculty of marginal quality or individuals seeking an entry position into academe as an alternative to their current career. These individuals have had either tenuous or no connections to the academic job market. We posit that the applicant pool to a non-tenure institution may have a relatively high proportion of unqualified applicants simply due to an increased proclivity to apply on their part.

Attraction and Retention of Faculty - Institutional Perspective

[2] . To measure this phenomenon, it is critical to know what other job positions were actually offered to the individual.

The institution's utility (defined as the decision to hire/retain an individual faculty member), we posit the following utility function:

$$Utility = f(W, RR, TR, SR, MC, F)$$

Where: W = Wage
RR = Research Reputation or Potential
TR = Teaching Reputation or Potential
SR = Service Reputation or Potential
MC = Expected Future Market Conditions for the Institution
F = a vector of factors relating to the fit between the individual candidate and the institution

We again wish to focus on the tenure vs. non-tenure issue. As institutions seek to maximize utility, they must choose some mixture of tenure track slots and non-tenure track slots. As we noted in the previous section, this choice does not hinge upon entry wages because they are approximately equal for tenured and non-tenured slots. We posit that the key initial factor when deciding between tenure track slots and non-tenure track slots is the forecast of future market conditions. This forecast is based upon the expected supply of available of faculty and the expected demand for their services. The demand for services is a derived demand emanating from enrollment and can be proxied by FTEs.

Forecasting future market conditions in academe is difficult. Unfortunately, where forecasts are made, institutions often employ rather naïve forecast models. Many of these models are essentially linear interpolations of the enrollment patterns among traditional students adjusted for obvious demographic changes.

It is important to note that an efficient adjustment process does not characterize the market for faculty services. For example, Ph.D. granting institutions exhibit

relatively low supply elasticity. The investment period for earning a Ph.D. is long so the supply response is slow. In addition, it is in the interest of Ph.D. granting institutions, at least over the short -run, to produce more Ph.D.'s regardless of market conditions upon exit from the program. Thus, the production of Ph.D.'s in response to changes in demand is likely to be a rather long and protracted process. On the demand side faculty services, as proxied by FTE enrollment, are likely to be relatively inelastic. While FTEs drive the demand, there are numerous institutional and funding constraints which make demand moves sluggish. As a result, the market for faculty services exhibits an inability to adjust rapidly to changes in supply and demand. This can result in relatively prolonged situations where the labor market is considered “soft” or “tight”.

In order to make an optimal decision the institution must balance the value of the flexibility associated with using non-tenure granting slots against the costs of selecting a non-tenure granting strategy. These costs are impacted by four major factors: transaction costs, the risk premium embodied in wage, faculty mix (by rank), and the general conditions of the labor market relative to forecasts for faulty demand.

The transaction costs associated with faculty can be substantial but are unlikely to be prohibitive to an institution for a number of reasons. For the purposes of this analysis we can break transaction costs down into the following component parts: search, negotiation, and contract consummation. The costs of search are likely to be most substantial because negotiations are relatively short and contracts well-defined. However, many of the search costs are incurred by faculty rather than by the institution. Thus, we posit that transactions costs are relatively low from the institutions perspective. The same analysis holds true for tenure granting institutions and non-tenure granting

institutions. The major difference between the two types of institutions would be in the number of searches if non-tenure granting institutions exhibit higher turnover rates among faculty[3].

As outlined in the previous section, wages are impacted by the compensating differential for job security. In an institution that chooses a high non-tenure granting mix, the risk premium associated with accepting a position without tenure must be periodically renewed. This means the costs of renewing mobile faculty will be rise. Thus the cost differential of a non-tenure granting strategy relative to a tenure granting strategy is impacted by the rate of decline in the risk premium associated with the compensating differential for job security[4] and the institution's willingness to accept turnover of mobile faculty. This leads directly to questions of faculty mix.

For non-tenure granting institutions, they are more likely to have to mark renewal contact wages to market given they are transferring risk to the faculty member. This will be especially true for more mobile faculty. We suspect this group of faculty would be comprised primarily of professors at the associate rank[5] (Marking professors to market will increase the labor costs of the institution).

If an institution chooses to accept high turnover rates and the concomitant reputation effects, faculty mix will be impacted. As noted earlier, non-tenure granting institutions that develop a reputation for non-renewal, will face a labor market where

---

[3] Chait and Ford, 1982 find that "Contract systems do not produce significant faculty turnover as a result of nonreappointments." Their conclusion is drawn from a 1972 by the American Council on Education analyzed by El-Khawas and Furniss, 1974.

[4] One articles explain how the Boston University School of Business has adjusted to exactly this issue. Tenured faculty were given the option of maintaining their tenure status or moving to a 10-year contract. "For those who opt for the 10-year alternative, a salary premium of 8 to 10 per cent is paid from the first day of employment, offsetting any perceived risk of forgoing a lifetime guarantee." See *A Realistic Alternative to Tenure*, The Chronicle of Higher Education, June 26, 1998, Volume XLIV, No. 42: p.B6.

faculty will reduce their supply of services. This information is most likely to be more widely utilized at the higher ranks of the professorate, at least initially. Thus, attracting faculty at the associate and full ranks becomes more difficult than attracting assistant professor and instructor ranks, and the portfolio of faculty mix will become skewed toward the lower ranks.

Costs will be impacted by faulty mix for two reasons. First, a high proportion of assistant and instructor positions will reduce costs for the institution. Second, renewal in a non-tenured granting institution must include the risk premium being marked to market.

Of course, these effects will take time. Many institutions experimenting with non-tenure slots are offering them to new entrants only. They are, in some respects assuring themselves of a faculty that maintains a traditional mix of ranks until attrition changes the mix naturally. However, under current forecasts, this entropy process may be speeded by the aging of the professorate and the expected future demand for faculty. If these two factors are forecasted correctly this strategy will not result in a stable faculty mix over the long term.

The last major factor affecting costs is the labor market and expectations concerning it. If labor markets are expected to soften or to remain soft, the institution may realize increased value from the flexibility afforded by using non-tenure granting slots. If the value of this flexibility is greater than the increase in cost encountered with a non-tenure granting strategy, the institution will increase its mix of non-tenure granting slots.

---

[5] It is widely held among the professorate that mobility drops dramatically once the rank of full professor is attained.

If labor markets are expected to tighten or remain tight, the value of the flexibility afforded by using non-tenure granting slots would be minimal. The institution will anticipate increased competition in the market for faculty services as a result of decreased supply of faculty or increased enrollment. Unlike the situation where market conditions are expected to soften or remain soft, the value of the flexibility afforded by using non-tenure granting slots will not be large enough to offset the more rapidly declining compensating differential for security associated with tenure granting slots. Therefore, the institution can minimize costs by offering tenure granting slots when market conditions are expected to tighten or remain tight.

If an institution is unsure about future market conditions, it may offer a mix of tenure granting and non-tenure granting slots. Selection of a mixed strategy may also be the result of the institution's existing faculty composition. In either case, it must be noted that selection of a mixed strategy may give rise to significant costs. The strategy of using a mix of tenured and non-tenured slots for faculty with identical job descriptions may result in the development of a two-tiered evaluation system. The existence of such a system may result in a decline in the value arising from the flexibility of the mix of tenured and non-tenured slots. This decline in the value of the flexibility afforded by the mixed strategy would affect the institution's calculation of the cost-benefit relationship that was previously used in the strategy selection process.

Summary

We apply a human capital approach to the problem of attracting and retaining faculty in a non-tenure granting environment and presented utility functions for individual faculty and institutions. Our analysis indicates the primary difference between tenured and non-tenured slots is the level of job security. In the framework presented, the compensating differential for job security as it relates to tenure practices is shown to be the key element in the attraction and retention process.

Three institutional strategies for faculty mix exist. Each strategy can be optimal under certain market forecast. However, each strategy involves a degree of risk on the part of the institution. The risk faced by the institution is of an incorrect forecast of future labor market conditions or conflicts arising from a two-tiered evaluation system.

References

Baltimore, David. "In Defense of Yenure." *Technology Review*, 1990: 70-71.

Brown, Ralph S., and Jordan E. Kurland. "Academic Tenure and Academic Freedom." *Law and Contemporary Problems*, 1990: 325-355.

Chait, Richard P., and Andrew T. Ford. *Beyond Traditional Tenure.* San Francisco: Jossey-Bass, 1982.

Diamond, Robert M. "The Tough Task of Reforming the Faculty-Rewards System." *The Chronicle of Higher Education*, 1994: B1-B3.

Finkin, Matthew W. *The Case for Tenure.* Ithaca: ILR Press, 1996.
—. "Scrapping Tenure Would Raise Costs." *USA Today Magazine*, December 1996: 14.

George, Richard T. de. *Academic Freedom and Tenure: Ethical Issues.* Lanham: Rowman & Littlefield, 1997.

Ladenson, Robert F. "Is Academic Freedom Necessary?" *Law and Philosophy*, 1986: 28-87.

Lataif, Louis. "A Realistic Alternative to Tenure." *The Chronicle of Higher Education*, 1998: B6.

Machlup, Fritz. "In Defense of Academic Tenure." *AAUP Bulletin*, 1964: 112-124.

McGee, Robert W. "Academic Tenure: Should It Be Protected By Law." *Western State University Law Review*, 1993: 593-602.

Mooney, Carlyn J. "Tenured Faculty Members Are Spared in Latest Round of Belt Tightening." *The Chronicleof Higher Education*, 1993: A17-A18.

Professors, American Association of University. "2006-07 Report on the Economic Status of the Profession." 2007.

USA Today Magazine. "Financial Squeeze Forcing Change." December 1993: 10.

Whicker, Marica Lynn. "An Economic Perspective of Academic Tenure." *PS*, 1997: 21-25.

Wilson, Robin. "Contracts Replace the Tenure Track for a growing Number of Professors." *The Chronicle of Higher Education*, 1998: A12-A14.

Yarmolinsky, Adam. "Tenure: Permanence and Change." *Change*, 1996: 16-20.

Time Spent Online and Student Performance in Online Business Courses: A Multinomial Logit Analysis[1]

Damian S. Damianov*[2]*, Lori Kupczynski, Pablo Calafiore, Ekaterina P. Damianova, Gökçe Soydemir, and Edgar Gonzalez

Abstract

This study examines the determinants of academic achievement in online business courses. As a measure of effort, we use the total amount of time each student spent in the course. We estimate a multinomial logistic model to examine the odds of attaining one grade versus another depending on time spent online, GPA, and some demographic characteristics of students. Our findings suggest that extra effort can help a student move from letter grades F, D and C to grade B, but is less helpful for the move from B to A. For the latter improvement, a high GPA matters the most.

Introduction

The determinants of academic performance are a recurrent topic in public policy debates on higher education. One largely unsettled issue concerns the impact of the most essential factors in the educational production—student's effort and study time—on academic achievement. While many would probably agree that students will not learn unless they put forth some amount of effort, our understanding of the ways study time impacts performance as measured by attaining a certain course grade is rather limited. Quantifying the effect of study time on achievement seems important from at least three perspectives: from the perspective of the administrator who is in charge of the design of effective teaching policies; from the perspective of the instructor, who creates classroom learning experiences and measures learning outcomes; and finally from the perspective of the student who seeks to balance competing personal goals.

In recent years much effort has been dedicated to understanding the factors contributing to the success of undergraduate business students. Nonis, Philhours, Syamil, and Hudson (2005) analyze survey data containing demographic, behavioral, and personality variables of 228 undergraduate students attending a medium size AACSB accredited public university. Using a hierarchical regression model they find that self-reported time per credit hour spent on academic activities outside of class explains a significant portion of the variation in the semester grade point average (GPA) for senior students, but has no impact on the cumulative GPA. Brookshire and Palocsay (2005) analyze the performance of undergraduate students in management science courses and report that overall academic achievement as measured by students' GPA has a significantly higher impact on achievement than students' mathematical skills as measured by math SAT scores.

Most of the existing literature on the topic relies on surveys in which students self-report the amount of time spent in a particular course or in a particular time frame, which is usually a semester or a year (see, e.g. Michaels and Miethe, 1989; Borg, Mason and Shapiro, 1989; Park and Kerr, 1990; Didia and Hasnat, 1998; Nofsinger and Petry, 1999; Cheo 2003; Williams and Clark 2004). The major concern associated with this approach is the accuracy and completeness of the collected data. Stinebrickner and Stinebrickner (2004) emphasize that the reporting error from retrospective survey questions is likely to be substantial. They discuss estimators that might be appropriate when reporting errors are common yet highlight the limitations of the results obtained from the analysis of such data samples.

[1] We thank José A. Pagán, Marie T. Mora, Alberto Dávila, Teofilo Ozuna, and Dave Jackson for insightful discussions. We are also thankful to two anonymous reviewers for many constructive suggestions. We are further grateful to the participants at the 36th annual meeting of the Academy of Economics and Finance in Pensacola, FL for their comments.

[2] Corresponding Author. Department of Economics and Finance, University of Texas—Pan American, 1201 West University Drive, Edinburg, TX 78539, Tel. (956) 533-9305, Fax (956) 384-5020, email: ddamianov@utpa.edu

This study, in contrast, analyzes *actual* rather than *self-reported* data on student activities in *online* courses during the course of an entire semester. Our sample includes 532 students who were enrolled in 13 courses offered online by the College of Business at a large public university in South Texas in the spring semester of 2008. We merge detailed information on student activities in online courses (in particular the total amount of time spent online) with administrative data on demographic characteristics and academic performance prior to taking these courses. The online course tracking device we use keeps a detailed record of individual student activity with a precise measure of the time each student spends on each activity of a course.[3]

The role of effort as measured by time spent on academic activities has been the focus of much research in recent years. A major limitation of this research is the paucity of actual (i.e. not self-reported in surveys) data on student activities. Johnson, Joyce and Sen (2002) measure student effort by the number of attempts and the amount of time spent by students on computerized quizzes. Lin and Chen (2006) estimate the relationship between attendance and exam scores whereby they differentiate between cumulative class attendance and attendance of lectures. A similar approach is taken by Romer (1993), who advocates for mandatory attendance based on the results from an extensive study performed on a data sample from three schools: a medium-sized private university, a large public university, and a small liberal arts college. Rich (2006) analyzes a self-collected data sample of students who took his senior-level corporate finance class at Baylor University. He finds a positive and significant relationship between grades and several measures of effort including class attendance, arriving on time, contribution to class discussion, and attempts at homework problems. All these analyses are based on ordinary least squares (OLS) regressions.

The present study differs from this literature in its empirical strategy. We use a multinomial logistic model (MNLM) with five different categories (grades A, B, C, D, and F) rather than an OLS estimation. Thus, our methodology will be similar to the one adopted by Park and Kerr (1990) but our data would be actual rather than self-reported. This methodology enriches previous analyses in two distinct ways. First, unlike the OLS regression, the MNLM is more appropriate for discrete dependent variables such as a course grade. Spector and Mezzo (1980), among others, contain a discussion of the inadequacies of the OLS method. In particular, they discuss the assumptions underlying OLS that will be violated when the dependent variable is not continuous. Second, the MNLM renders valuable additional information that is not obtainable using OLS. An OLS coefficient provides the marginal effect of, say, one additional hour of study on the grade attained. This coefficient does not differentiate between the additional effort of a student necessary for moving from C to B and the one necessary for moving from B to A. The MNLM, in contrast, provides this additional information, and, as we will see, it is substantially more costly for a student to move from B to A rather than from C to B as far as time spent online is concerned.

The rest of the paper is organized as follows. Section 2 contains a description of our data sample and section 3 introduces the multinomial logit model. Section 4 presents the empirical results, section 5 discusses alternative model specifications, and section 6 concludes.

Data description

We obtain data on students enrolled in online business courses during the Spring 2008 semester at the College of Business of a large public university in South Texas from two sources. The first source contains the track record of student activities in the online courses. We focus only on fully online courses offered by the College of Business. Students in these courses do not necessarily meet face-to-face with the instructor throughout the semester. All the course instructors who taught these courses are fulltime tenure track or tenured professors. The university uses Blackboard (campus edition) as a Learning Management System and keeps detailed student activity data per class in Blackboard's tracking record summary. The second data source contains demographic and academic information about students obtained from the University's Office of Admissions and Records. We merge both databases and eliminate any identifier that could lead to the recognition of an individual subject in order to ensure anonymity.

[3] This university uses Blackboard (Campus edition) as a Learning Management System for online courses. This software allows instructors and students to meet in a closed area online to participate in coursework. Blackboard's activity tracking feature keeps a record of various student activities, including log in and log out time of each session, number of sessions, breakdown of the time spent in various learning modules, participation in discussion boards (including messages posted and messages read), number of emails sent and received for each student, etc. For the period of the Spring semester of 2008, the vast majority of the students who took an online class (89.1%) used a high-speed internet connection such as cable-modem, T1 or broadband to gain access to the courses. About 2.1% used a dial up connection, and the remaining 8.8% used other types of connections (figures retrieved from Google Analytics for all users accessing the university's online learning web site during the Spring semester of 2008).

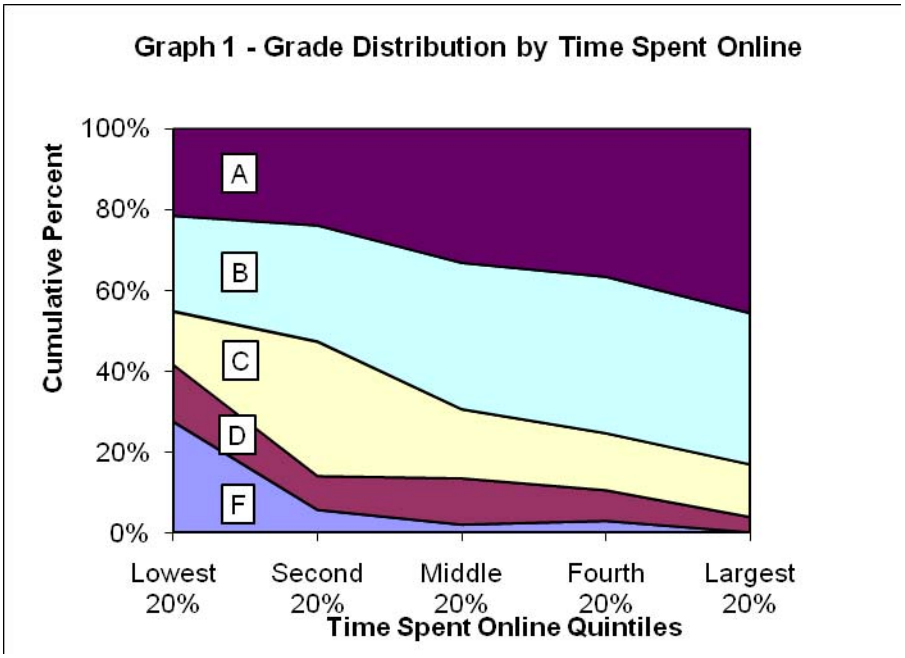
From the initial sample of 563 students we remove those that voluntarily dropped the course before the 12th day of classes. The final sample consists of 532 students who enrolled and completed one of 13 online courses offered by the four departments in the College of Business: Economics and Finance (Introduction to Economics, two sections of Principles of Microeconomics I, Managerial Finance, and International Finance); Accounting and Business Law (Professional Ethics, and Business Law I); Computer Information Systems (Management Information Systems); and Management, Marketing and International Business (Communication Policy, Principles of Marketing, International Marketing, and two sections of International Business).

Table 1 shows that, on average, almost 41 students were enrolled per online class. The average student age was 25 years and the median age was 23 years. There were 215 (40.4%) male students and 317 (59.6%) female students. Given that our sample comes from a minority serving institution, more than 85% of the students were of Hispanic origin. There were 431 full-time and 101 part-time students enrolled. The majority of students described themselves as Senior (56.2%) followed by Junior (28.9%).

Table 1: Descriptive statistics of 532 students enrolled in an online business course during the Spring 2008 semester

Class characteristic	#	**%**
Average class size (number of students)	40.92	
Average student age (in years)	25.07	
Median student age (in years)	23.00	
Male (number of students)	215	40.4%
Female (number of students)	317	59.6%
Hispanic (number of students)	454	85.3%
Non Hispanic (number of students)	78	14.7%
Full-time students	431	
Part-time students	101	
Freshman (number of students)	12	2.3%
Sophomore (number of students)	37	7.0%
Junior (number of students)	154	28.9%
Senior (number of students)	299	56.2%
Graduate (number of students)	30	5.6%
Average student GPA	2.86	
Median student GPA	2.84	
Variance in student GPA	0.25	
Grade distribution		
F (number of students)	40	7.5%
D (number of students)	48	9.0%
C (number of students)	97	18.2%
B (number of students)	175	32.9%
A (number of students)	172	32.3%

The average student grade point average (GPA) prior to taking the online class was 2.86 while the average and the median grade obtained per enrolled student was a B, with a variance of 1.48. At the end of the semester, the number of students who received grades A, B, C, D, and F were 172 (32.3%), 175 (32.9%), 97 (18.2%), 48 (9%), and 40 (7.5%), respectively. Graph 1 shows the final grade distribution depending on the time students spent online in the class. In some courses, the average time spent per student was significantly higher than in others. In order to make time comparable across classes, in each course we standardize the time students spent online. From the actual time (in minutes) for each student we subtract the average time for all students in this course and divide the result by the standard deviation of time spent in the course. Standardized time is thus normally distributed with a mean of zero and a standard deviation of one. Blackboard automatically logs students off from courses after 20 minutes of inactivity. Therefore, we believe that our measurement captures quite accurately the actual time students spent online. Graph 1 illustrates that students who spent the least amount of time online (those included in the lowest quintile) earned the largest proportion of failing grades (F). In contrast, the students who spent the most time online received the smallest proportion of F grades. This group also received the largest proportion of A grades. Generally, Graph 1 suggests a positive relationship between time spent online and final grade.



Multinomial logit model (MNLM)

We use a multinomial logistic regression to analyze how the odds of receiving a specific grade versus another depend on the time the student spent online and the student's GPA prior to taking the course. These two variables, which are broadly interpreted as student *effort* (or preparation) and *ability* (or intelligence), are major inputs in the learning production function. We also control for age, gender, major, and number of credit hours taken prior to enrolling in the course. Thus, our model has six explanatory variables defined as follows:

GRADE = the letter grade (A, B, C, D, F) the student received in the course, with A = 4, B = 3, C = 2, D=1, and F = 0

GPA = the grade point average (scale 0-4.0) of the student at the beginning of the semester.

TIME = the actual time the student spent online working on course content, standardized per class.

AGE = age of the student at the time he/she took the course.

GEN = 1 if student is female; 0 if male.

MAJOR = 1 if the online class is offered by the same department of the student's declared major; and 0 otherwise.

PHRS = the total credit hours the student had accumulated prior to enrolling into the course.

As a base category (or comparison group) we choose the category with the largest number of observations (see Long 2006, p. 231). In our case the most frequently assigned grade is the letter grade B. More specifically, the MNLM specifies the logarithm of the odds of grade $i = A, C, D, F$ versus grade B as a linear function of the explanatory variables. Thus, the MNLM is specified by four equations:

$$\ln\left(\frac{\Pr(i)}{\Pr(B)}\right) = \alpha_{i|B} + \beta_{1,i|B} GPA + \beta_{2,i|B} TIME + \beta_{3,i|B} AGE + \beta_{4,i|B} GEN + \beta_{5,i|B} MAJOR + \beta_{6,i|B} PHRS$$

where $i = A, C, D, F$. These four equations can be solved to calculate the probabilities for each grade i as a function of the six explanatory variables (and the regression estimates for the coefficients):

$$\Pr(i) = \frac{\exp(Z\beta_{i|B})}{\sum_{j=A,B,C,D,F} \exp(Z\beta_{j|B})}$$

where $Z = (GPA, TIME, AGE, GEN, MAJOR, PHRS)$ is the vector of explanatory variables and $\beta_{j|B} = (\beta_{1,j|B}, \beta_{2,j|B}, \ldots, \beta_{6,j|B})$ is the vector of coefficients.[4] The coefficients are obtained using maximum likelihood estimation, which ensures consistency of the estimates when explanatory variables are categorical (see e.g. Park and Kerr, 1990). The estimation is based on the *independence of irrelevant alternatives* assumption, which implies that the odds of one grade versus another do not depend on the availability of other grades. We confirm the validity of this assumption for our data sample by performing a Small-Hsiao test (see Long 2006, p. 245). The results of this test are reported in Table A.3 in the Appendix. A direct check of the correlation coefficients (see Table A1 in the Appendix) and some preliminary OLS estimations also reveals that no multicollinearity exists between explanatory variables in our data sample.

Empirical results

Our major findings are presented in Table 2, which contains the multinomial logit coefficients for the logarithms of the odds of all grades versus the base category B, followed by the z-values in parenthesis, and the marginal effects of each explanatory variable on the probability of receiving a particular grade. The marginal effects are calculated by the formula

$$\frac{\partial \Pr(i)}{\partial x_k} = \Pr(i)\left[\beta_{k,i|B} - \sum_{j=A,B,C,D,F} \beta_{k,j|B} \Pr(i)\right]$$

where x_k, $k = 1, 2, \ldots, 6$ denote the explanatory variables GPA, TIME, AGE, GEN, MAJOR, and PHRS, respectively. The partial derivative is evaluated at the sample mean of each regressor. In contrast to the *multinomial logit coefficients*, which specify the impact of each explanatory variable on the log-odds ratio of one grade versus another, the *marginal effects* determine the impact of a small change in each explanatory variable directly on the

[4] The coefficients $\alpha_{i|B}$ are normalized to zero to ensure the identification of the coefficients in the $\beta_{j|B}$ vectors (see e.g. Greene, p. 721).

probability of receiving a particular grade. As is evident from Table 2, the marginal effect of TIME is positive for grades A and B, and negative for grades C, D and F. The marginal effect of GPA is positive only for grade A and negative for all other grades.

Table 2: Multinomial Logit Estimates and Marginal Effects of the Time Spent Online on Class Grade

	Grade B	Grade A	Grade C	Grade D	Grade F
Constant	–	-5.803***	4.172***	8.125***	3.872**
		(-5.38)	(3.52)	(4.96)	(2.12)
GPA		1.558***	-1.243***	-3.029***	-2.155***
		(5.79)	(-3.7)	(-5.65)	(-3.89)
	-0.022	0.464	-0.266	-0.146	-0.031
TIME		0.061	-0.362**	-0.741***	-2.751***
		(0.52)	(-2.34)	(-3.11)	(-6.21)
	0.055	0.058	-0.048	-0.029	-0.036
AGE		0.019	-0.020	-0.027	-0.000
		(0.97)	(-0.81)	(-0.79)	(0.00)
	-1.714E-04	0.006	-0.004	-0.001	-0.000
GEN		-0.022	-0.226	0.284	-0.162
		(-0.09)	(-0.85)	(0.77)	(-0.39)
	0.018	0.006	-0.038	0.015	-0.002
MAJOR		-0.068	-0.166	-0.598	0.214
		(-0.28)	(-0.58)	(-1.41)	(0.49)
	0.034	0.003	-0.018	-0.023	0.004
PHRS		0.006	-0.006	-0.009	-0.011
		(1.33)	(-1.26)	(-1.34)	(-1.31)
	4.150E-05	0.002	-0.001	-0.000	-0.000

Log likelihood: -630.46

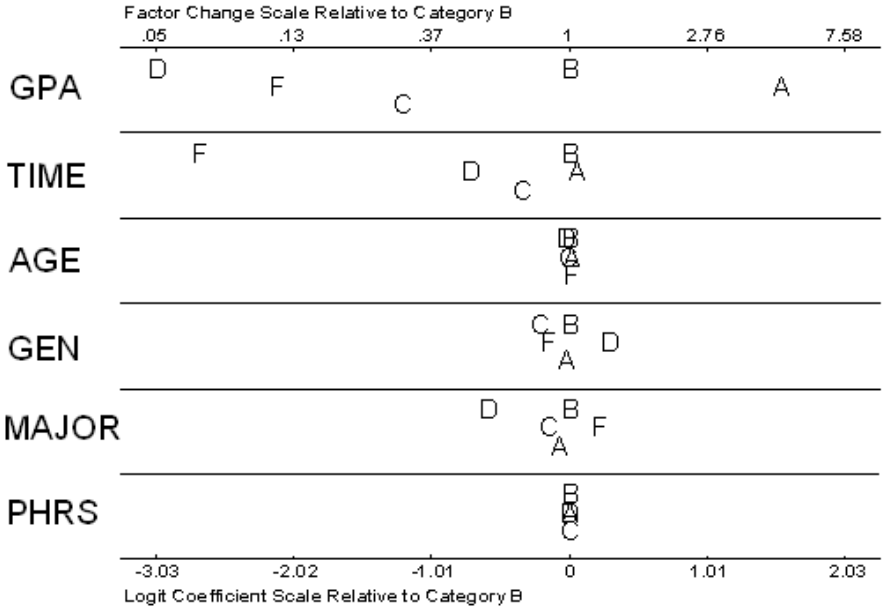
Pseudo R^2 : .181

N = 532

Notes: The logit coefficients are followed by the *z*-values (in parenthesis) and the marginal effects evaluated at the sample means. The reference category is Grade B. ** and *** indicate significance at the 5% and 1% level respectively, using two-tailed tests.

The GPA coefficients for the logarithm of the odds for all grades versus B are significant at the 1% level and have the expected signs since letter grade B is the reference category. The TIME coefficients are also significant at the 1% level except for the coefficients of the odds A vs. B (non-significant), and the odds of C vs. B (significant at the 5% level). For instance, a unit increase in the time spent online increases the log odds of getting an A vs. B by 0.061 and decreases the odds of getting a C vs. B by 0.362. Note also that the coefficients for TIME monotonically increase moving from F upward which is consistent with diminishing marginal returns to effort. The coefficients for all other variables are insignificant. This may suggest that there may be an identification problem inherent to our model specification. The goodness of fit of our logistic regression as measured by the Pseudo R^2 is 0.181, which may imply the possibility of other variables having an impact on the log-odds of one grade versus another.

Graph 2: Odds-ratio plot of grade relative to base category B



These results are also visualized in Graph 2. Each row in this graph presents the effect of an independent variable on the log odds ratios (scale is above the graph) of each grade versus grade B. If a letter grade is positioned to the right of another letter grade, then an increase of the explanatory variable makes the odds for the outcome to the right more likely (Long 2006). In fact, the distance between each pair of letters indicates the magnitude of the effect because it corresponds to the value of the logit coefficients reported in Table 2. As is evident, the impact of GPA on grade is almost evenly spaced across grades, with letter grade A to the very right, followed by B and C. The spaces A – B and B – C seem equally large and a little larger than the spaces between the other grades, indicating that ability plays an important role in determining a student's chance of receiving grades A or B. The ordering of D and F is somewhat surprising because it indicates that a high GPA makes it more likely for a student to receive an F instead of a D. One possible explanation for this observation is that students with a good GPA prefer to fail and retake a class in which they see that they will obtain the lowest passing grade.

Two issues are worth emphasizing when it comes to the effect of TIME. First, the failing grade (F) lies much further to the left of all other grades. This indicates that a very small increase in student time spent online will substantially reduce a student's odds of failing a course. Second, the grades A and B are clustered together (although the ordering is as expected). This signals that the odds of getting an A versus B cannot be improved upon substantially by increased amount of time spent online. Ability seems much more important than effort for an A grade in the courses. The letter grades for all other variables are clustered together, which indicates that the impact of these variables is minimal.

Table 3: Odds of getting one grade versus the next best grade

	A vs B	B vs C	C vs D	D vs F
GPA	1.557***	1.242***	1.785***	-0.873
	(0.269)	(0.335)	(0.541)	(0.648)
TIME	0.061	0.361**	0.379	2.009***
	(0.117)	(0.154)	(0.246)	(0.461)

Notes: the coefficient values are followed by standard errors in parenthesis. ** and *** indicate significance at the 5% and 1% level respectively, using two-tailed tests.

A common student question concerns the amount of effort necessary for a student to move to the next best grade, and Table 3 provides a perspective on this issue. It contains the coefficients for GPA and TIME for the odds of moving to the next best grade. As is evident, previous track record of high performance as measured by student's GPA will improve the odds of receiving a higher grade. At the conventional significance levels, the odds of getting one grade versus the next best grade are all statistically significant with the exception of the last column which reports the odds of moving from F to D. For this move, effort is much more crucial. Effort will also help a student move from C to B (the coefficient is significant at the 5% level).

The estimated model can predict the expected achievement of a student with certain characteristics based on the effort of the student as measured by the time spent online. Table 4 presents the probabilities for receiving various grades of a 23-year-old female student taking a course in the department of her major with a current GPA of 3.0, and 18 credit hours already taken. The first row presents the probabilities for obtaining various grades if she spends a time equal to the average time (AVG) that students spend in class, and the second row reports the probabilities in case she increases her time spent online to one standard deviation above the average (AVG+1SD).

Table 4: Impact of time on the chances of receiving a particular grade for a student with certain characteristics.

	A	B	C	D	F
AVG	21.95%	43.62%	26.00%	5.60%	2.83%
AVG+1SD	26.53%	49.63%	20.60%	3.04%	2.10%

Table 4 reveals that, as a result of the extra effort, the chance of this student of receiving an A increases by 4.48% and the chance of receiving a B by 6.1%. The chances of receiving each one of the other grades C, D, and F decreases.

The MNLM assumes that the logarithms of the odds of one grade versus another are linear in the independent variables. To test the validity of this assumption we also considered two alternative specifications of the model. To check for nonlinear relationships regarding time spent online we added a squared term of this variable; and to test for possible interaction between the ability of the student (as measured by the GPA of the student) and effort (as measured by time spent online) we considered a regression with an interaction term of these two variables. The coefficients for both the interaction term and the squared term are not significant, which lends support for the final

specification adopted. As an alternative, we also analyzed subsamples of the data. We created the following subsamples: economics and finance and computer information systems (EF+CIS); computer information systems and marketing, management and international business (CIS+MMIB); and marketing, management and international business and economics and finance (MMIB+EF).[5] The results for the subsamples are found to be qualitatively the same as the ones for the entire database.

Discussion

This study focuses on the role of time spent online on academic performance. While we believe the total amount of time spent on a course is a good measure of student effort, it might be interesting to further explore how this time is allocated to various activities. Alternative measures of effort in our online courses can be number of sessions (i.e., number of times a student logged into the course), number of emails, number of messages posted on discussion boards, etc. An analysis of those measures might shed further light on the ways students learn, and which activities, on average, contribute to performance.

Our study is based on data collected from online courses in one large public university, but as data from online courses become available it will be important to expand the current study to data from courses in other institutions. It will also be interesting to compare the effect of time between online and traditional courses yet the progress in this direction will clearly depend on the availability of reliable data from traditional courses.

Sometimes students are content getting a particular grade and seek to dedicate the minimum effort possible which will guarantee that they will attain this grade. Thus, the factors which motivate students to expend effort might be different depending on whether students are grade satisfiers or grade maximizers. Our dataset is static in nature and does not allow us to differentiate between the two groups of students or somehow classify students depending on their motives to spend time studying. To this end we need to have detailed information on student performance throughout the semester.[6]

Conclusion

The literature on the impact of student effort on performance naturally divides in two categories based on the data used and the empirical results. The analyses relying on *self-reported* measures of student effort in general tend to find a weaker or no relationship between effort and performance, while the analyses using *real* measures of effort find a much stronger link between effort and educational outcomes. For instance, Stinebrickner and Stinebrickner (2003) find that working students consistently attain a lower GPA compared to full-time students, and Dávila and Mora (2004) find that students with entrepreneurial parents have a lower achievement in mathematics and reading compared to students coming from salaried households.

This paper uses data on the real time students spent in online classes to examine the link between the time put forth in a particular online class and the grade attained in this class. We find a positive and significant relationship between study time and grade; yet, our analysis also uncovers the limits of what a student can achieve by putting forth extra effort in a particular course. While study time is quite helpful for a student to move away from the grades F, D and C, and attain a B instead, effort, although beneficial, has only a statistically insignificant contribution for the improvement from B to A. For this improvement, overall GPA is the more important determinant.

The number of institutions offering online courses has grown steadily in the last several years and student enrollment in online courses has also risen significantly. Therefore, understanding the determinants of academic success in online classes has become increasingly important to administrators, professors, and students. Our findings suggest that activities encouraging students to spend more time online will result in a higher educational achievement. Students should know that educational success will be much more likely when time and effort is put forth from the day they come to the university.

---

[5] We would like to thank an anonymous referee for this suggestion.

[6] We thank an anonymous reviewer for pointing out this issue.

References

Borg, Mary O., Mason, Paul M. and Shapiro, Stephen L. 1989. "The Case of Effort Variables in Student Performance." *Journal of Economic Education* 20: 308-313.

Brookshire, Robert G. and Palocsay, Susan W. 2005. "Factors Contributing to the Success of Undergraduate Business Students in Management Science Courses." *Decision Sciences Journal of Innovative Education* 3: 99-108.

Cheo, Roland. 2003. "Making the Grade through Class Effort Alone." *Economic Papers* 22: 55-65.

Dávila, Alberto and Mora Marie T. 2004 "The Scholastic Progress of Students with Entrepreneurial Parents." *Economics of Education Review* 23: 287-29.

Didia, Dal and Hasnat, Baban. 1998. "The Determinants of Performance in the University Introductory Finance Course." *Financial Practice and Education* 8: 102-107.

Greene, William H. 2003. *Econometric Analysis*. New York: Prentice Hall, 5th ed.

Johnson, Dean L., Joyce, B. Patrick. and Sen, Swapan. 2002. "An Analysis of Student Effort and Performance in the Finance Principles Course." *Journal of Applied Finance: Theory, Practice, Education* 12: 67-72.

Lin, Tsui-Fang and Chen Jennjou. 2006. "Cumulative Class Attendance and Exam Performance." *Applied Economics Letters* 13: 937-942.

Long, J. Scott and Freese, Jeremy. 2006. *Regression Models for Categorical Dependent Variables Using Stata*, College Station: Stata Press.

Michaels, James W. and Miethe, Terance D. 1989. "Academic Effort and College Grades." *Social Forces* 68: 309-319.

Nofsinger, John R. and Petry, George R. 1999. "Student Study Behavior and Performance in Principles of Finance." *Journal of Financial Education* 25: 33-41.

Nonis, Sarath. A., Philhours, Melody, Syamil, Ahma D., and Hudson, Gail I. 2005. "The Impact of Non-Intellectual Variables on the Academic Success of Business Students." *Marketing Education Review* 15: 51-63.

Park, Kang H. and Kerr, Peter M. 1990. "Determinants of Academic Performance: a Multinomial Logit Approach." *Journal of Economic Education* 21: 101-111.

Romer, David. 1993. "Do Students Go to Class? Should They?" *Journal of Economic Perspectives* 7: 167-174.

Rich, Steven P. 2006. "Student Performance: Does Effort Matter?" *Journal of Applied Finance* 16: 120-133.

Stinebrickner, Ralph and Stinebrickber, Todd R. 2003. "Working During School and Academic Performance." *Journal of Labor Economics* 21: 449-472.

Stinebrickner, Ralph and Stinebrickber, Todd R. 2004. "Time-Use and College Outcomes." *Journal of Econometrics* 121: 243-269.

Spector, Lee C. and Mazzeo, Michael. 1980. "Probit Analysis and Economic Education." *Journal of Economic Education* 11: 37-44.

Williams, Robert and Clark, Lloyd. 2004. "College Students' Ratings of Student Effort, Student Ability and Teacher Input as Correlates of Student Performance on Multiple-Choice Exams." *Educational Research* 46: 229-239.

Appendix

Table A.1: Pairwise correlation coefficients

	GRADE	GPA	TIME	AGE	GEN	MAJOR	PHRS
GRADE	1						
GPA	0.519	1					
TIME	0.329	0.171	1				
AGE	0.081	-0.009	0.193	1			
GEN	0.097	0.104	0.113	0.074	1		
MAJOR	0.106	0.189	-0.052	-0.037	-0.001	1	
PHRS	0.229	0.196	0.071	0.139	0.029	0.257	1

Table A.2: Likelihood-ratio tests for independent variables
Ho: All coefficients associated with given variables are 0

	Chi Square	df	P>Chi Sq.
GPA	141.002	4	0.000
TIME	71.944	4	0.000
AGE	2.973	4	0.562
GEN	2.159	4	0.706
MAJOR	3.108	4	0.540
PHRS	7.369	4	0.118

Table A.3: Small-Hsiao test of the independence of irrelevant alternatives.
Ho: Odds (Outcome-J vs Outcome-K) are independent of other alternatives.

Omitted	lnL(full)	lnL(omit)	Chi Sq.	df	P>Chi Sq.	Evidence
Grade A	-180.974	-172.192	17.563	21	0.676	for Ho
Grade C	-201.379	-194.789	13.180	21	0.902	for Ho
Grade D	-244.244	-235.584	17.320	21	0.692	for Ho
Grade F	-271.994	-264.723	14.541	21	0.845	for Ho

Table A.4: Predicted probabilities

Variable	Observations	Mean	Std. Dev.	Min	Max
GPA	532	2.859	0.503	1.605	4.000
GRADE	532	2.735	1.215	0.000	4.000
Grade A	532	0.321	0.222	0.002	0.890
Grade B	532	0.329	0.107	0.025	0.590
Grade C	532	0.183	0.099	0.001	0.438
Grade D	532	0.091	0.106	0.000	0.612
Grade F	532	0.076	0.131	0.000	0.776

Parsimonious Expected Utility and In-the-Large Risk Premiums for the Undergraduate Curriculum

Richard Robinson[1]

Abstract

Expected utility theory, Pratt and Arrow's risk premium, and the Markowitz risk premium are based upon Von Neumann and Morgenstern's five axioms of cardinal utility. The derivation of these measures from these axioms is generally considered too mathematically sophisticated for the undergraduate curriculum. This leaves a wide gap in the foundation of financial theory for the undergraduate. An alternative simplified presentation of expected utility that is suitable for the undergraduate curriculum, and that avoids the nuances and complexities of the traditional mathematically more complete derivation, is presented here. In addition, an easily derived in-the-large risk premium is shown to be the product of a simple geometric measure of concavity and the standard deviation of the gamble. An application to futures markets is presented as an illustration. Keywords: Expected Utility, Risk Premium

Introduction

Risk premium analysis lies at the core of the financial economic theory involving choice under uncertainty. As such, the notion of the risk premium plays a paramount role in the finance pedagogy. For example, risk premium analysis is essential for explaining individual preferences for the expected value versus risk tradeoff. It is also essential for explaining the behavior of certain speculative markets. This analysis is universally present in financial theory texts, but these are generally relegated to the graduate curriculum. The typical undergraduate is assumed to lack the mathematical sophistication (or education) required to benefit from this elucidation.

In particular, risk premium analysis is viewed as relying upon expected utility theory with its five axioms of choice as originally presented by Von Neumann and Morgenstern (1948), Friedman and Savage (1948), and Herstein and Milnor (1953). These axioms concern comparisons of complex gambles, and as a result, the derivation of the expected utility hypothesis is rather advanced for an undergraduate finance course. Following Markowitz (1959), "in-the-large" risk premiums are predicated upon expected utility. Following Pratt (1964) and Arrow (1971), "in-the-small" risk premiums are derived from Taylor series expansions of expected utility functions. Of course, knowledge of series expansions is not expected of the undergraduate business major who usually is required to have only a single semester of calculus.

A simpler approach to expected utility, one that avoids the complexities involved with the independence and ranking axioms, is suggested here as being appropriate for the undergraduate curriculum. It is suggested that only two simple axioms of choice are sufficient for the undergraduate elucidation. The more rigorous analysis associated with Von Neumann and Morgenstern's axioms can be left for graduate course exploration.

A parsimonious approach to derivation of the in-the-large Markowitz risk premium is also presented here, one strictly involving only simple 50% - 50% gambles of monetary wealth. It is shown to be directly related to an easily understood geometric measure of the concavity of the utility function for wealth. Following this derivation, an application that illustrates the activities of hedgers and speculators in futures markets, one that explains the dominance of hedgers over speculators, is offered as appropriate for the undergraduate curriculum.

The entire presentation requires only basic algebra and two dimensional geometry with no calculus. It is therefore entirely suitable for the undergraduate finance curriculum. It is a hope that this parsimonious presentation will allow the undergraduate finance major to have a more complete foundation in risk theory.

[1] Professor of Finance, SUNY at Fredonia, E336 Thompson Hall, Fredonia, NY 14063, Richard.Robinson@fredonia.edu.

The Traditional Axioms of Choice Under Uncertainty, and Risk Premium Measures

Graduate level text presentations of choices involving risk generally begin with the *five axioms of cardinal utility*, first presented by Von Neumann and Morgenstern (1948), and reviewed below. For these presentations, the standard symbology of "$x \prec y$" indicates that "the individual prefers y to x," and "$x \sim y$" indicates that "the individual is indifferent between x and y." Also, "$G(x,y:\alpha)$ indicates a binomial gamble of receiving x with probability α, and y with probability $(1-\alpha)$." The five axioms are:

1. The *comparability* or *completeness axiom*: If S is the set of uncertain outcomes, then for any $x, y \in S$, the individual can state either $x \prec y$, or $y \prec x$, or $x \sim y$.

2. The *transitivity axiom*: If $x, y, z \in S$, then if $x \succ y$, and $y \succ z$, then $x \succ z$. Also if $x \sim y$, and $y \sim z$, then $x \sim z$.

3. The *strong independence axiom*: If $x \sim y$, then $G(x,z:\alpha) \sim G(y,z:\alpha)$.

4. The *measurability axiom*: If $x \succ y \succ z$, then there is a unique α, where $0 < \alpha < 1$ such that $y \sim G(x,z:\alpha)$.

5. The *rankings axiom*: If $x, y, z, u \in S$, and $x \succ y \succ z$, and $x \succ u \succ z$, and if $y \sim G(x,z:\alpha_1)$, and $u \sim G(x,z:\alpha_2)$, then it follows that if $\alpha_1 > \alpha_2$, then $y \succ u$, or if $\alpha_1 = \alpha_2$, then $u \sim y$.

From these five axioms, the expected utility function, as expressed by equation (1), is derivable.[2]

$$U\{G(x,z:\alpha)\} = \alpha U(x) + (1- \alpha)U(z) \qquad (1)$$

Axioms 1 and 2 above are comparable to the axioms of choice under conditions of certainty. We require that the individual be able to rank preferences for all possible goods, and that these rankings be rational or transitive. The

[2] Copeland and Weston (1983, p. 79-80) present one of many advanced text derivations of the expected utility hypothesis.

other 3 axioms of choice under uncertainty involve the comparability of multiple gambles involving complex combinations of goods.[3] For the purpose of presentation to undergraduate courses, it is sufficient to point out here that axioms concerning comparability and transitivity (rationality) are suitable, but those concerning independence, and cardinal rankings, such as axioms 3, 4 and 5 listed above, are probably not suitable. All five axioms are necessary for a full and rigorous investigation, but not for an initial undergraduate review of risk analysis. To accomplish this initial view, the traditional axioms are modified below.

In addition to the complexity involved with establishing the existence and derivation of the expected utility function, risk premiums are generally derived using Pratt and Arrow's method of equating the expected utility of the gamble to the utility of current wealth minus a risk premium to be paid to avoid the gamble and leave the individual indifferent. For an actuarially neutral gamble involving a random outcome of Z (where $E(Z) = 0$), and initial wealth of W_0, then for a risk premium of R, and utility of wealth function U, equation (2) holds where $R > 0$ for the risk averse individual.

$$U(W_0 + Z) = U(W_0 - R) \tag{2}$$

Using a Taylor series expansion of (2), Pratt (1964) and Arrow (1971) derived their version of the in-the-small risk premium as given by approximation (3), where U' is the first derivative of U with respect to wealth W, U'' is the second derivative, and σ^2 is the variance of the underlying probability distribution. Pratt (1964) first applied the term "in-the-small" for gambles of infinitesimal stakes, and "in-the-large" for finite stakes.

$$R \cong -\frac{1}{2}\sigma^2\{U''/U'\} \tag{3}$$

In (3), risk aversion requires that the ratio $-U''/U'$ be positive. This ratio was termed the "index of absolute risk aversion" by Pratt. It is a concavity measure for the function U. Following this tradition of Pratt and Arrow, any alternative and simpler measure of risk premium should also be related to a basic concavity index of the utility function. This is accomplished below through an obvious geometric index that should be easily understood by the undergraduate finance major.

Axioms for the Derivation of Expected Utility for the Undergraduate

For the undergraduate finance theory, it is suggested here that the gambles investigated should be only 50% - 50% binomial gambles that involve monetary wealth. Two reduced form axioms are necessary: the first is a combination of *comparability* and *transitivity*, and the second is a simplified compendium of *independence, measurability* and *ranking* in the sense that these three traditional axioms are imbedded in it. This second axiom is termed the *fair utility* axiom. This is certainly not a claim that the subtle arguments concerning the *independence axiom* are not valid, or that they are unimportant, but rather just that they are not suitable for undergraduate exploration. A logically consistent simplification of the traditional axioms is in order, and presented here. As with the five traditional axioms, it can be shown that expected utility is a necessary consequence of the two reduced form axioms presented below.

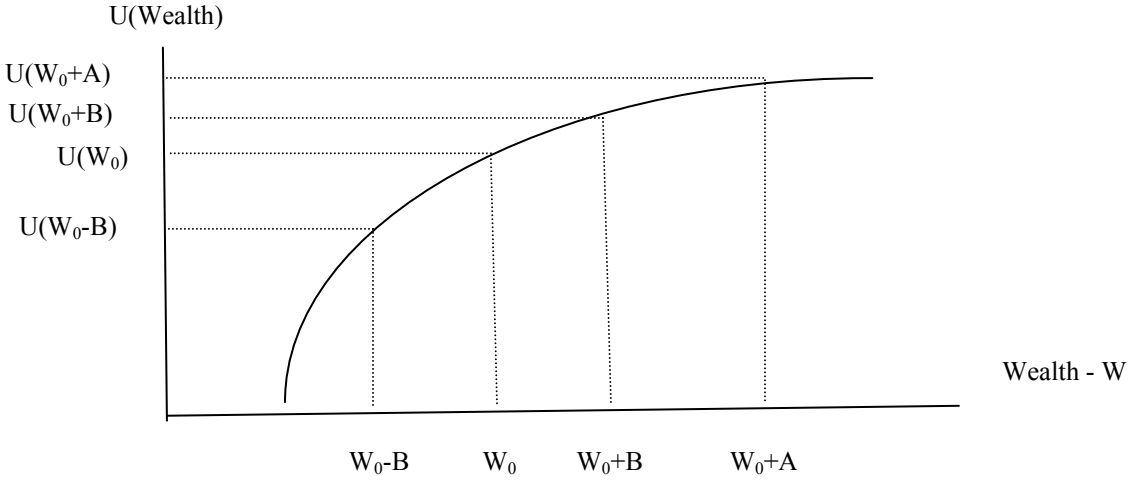
> *Axiom (1): More wealth is better than less*: For any levels of wealth $W_0 < W_1 < W_2$, then $U(W_2) > U(W_1) > U(W_0)$.
>
> *Axiom(2): Fair utility*: For a binomial gamble of 50% - 50% odds involving a wealth loss of B or win of A, where $A > 0$, and $B > 0$, then the individual is indifferent between taking the gamble or staying at initial wealth W_0 if the utility to be gained equals the utility to be lost as in (4).

3 Gambles such as a hotdog versus a hamburger as compared to a hotdog versus a tuna sandwich are complex and pose comparability problems involving the independence axiom. The expected utility literature, as developed over six decades, still debates the nuances of the *strong independence axiom*, and its variations. (See Machina (2009) as an example.)

$$U(W_0) - U(W_0 - B) = U(W_0 + A) - U(W_0) \quad (4)$$

Axiom 2 states that the gamble must be fair utility wise in order for indifference. This is illustrated by Figure 1. After Figure 1 is explored by the undergraduate, then *axiom* 2, together with the following definition of risk aversion, should be readily understood.

Figure 1
Sweetening the Winning Stakes by A - B



Definition of risk aversion: Axioms 1 and 2 imply risk aversion if A > B.

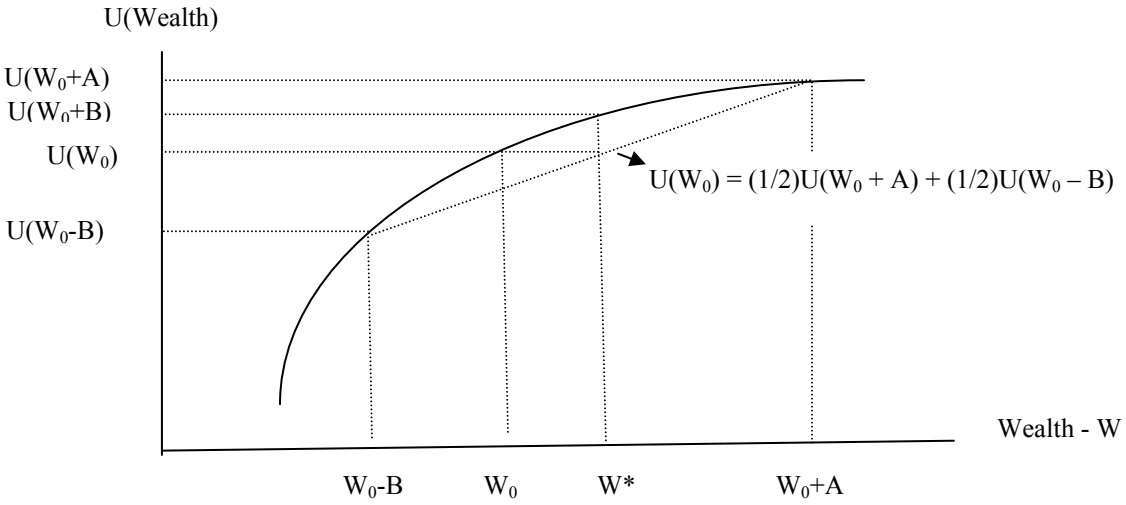
Once equation (4) is understood, as illustrated by Figure 1, then equation (5) algebraically follows: the expected utility of the 50% - 50% gamble for stakes A and B equals the utility of initial wealth, W_0, with certainty.

$$U(W_0) = (1/2)U(W_0 + A) + (1/2)U(W_0 - B) \quad (5)$$

Figure 2 illustrates the expected utility notion presented by equation (5).

Figure 2

Expected Utility and the Markowitz Risk Premium of W* - W_0



Note that (5) is a simple algebraic derivation from (4). The expected utility of the gamble follows directly from the *fair utility axiom* as expressed by equation (4). Expected utility is in fact embedded into *axiom 2*. This is the simplest derivation of expected utility possible, and it certainly is within the domain of the mathematical preparation of the undergraduate finance major.

Expected Utility and a Simple In-the-Large Risk Premium

For this undergraduate exposition, Figure 2 is adequate to illustrate the expected utility notion. It can also be used, however, to illustrate the Markowitz risk premium. Given the 50% - 50% odds, the expected wealth level W*, as given by (6), also lies at the expected utility point.

$$W^* = (1/2)(W_0 + A) + (1/2)(W_0 - B) \quad (6)$$

It is an elementary geometric exercise for the undergraduate to show that this intersection must occur as indicated in Figure 2.[4]

To derive the Markowitz in-the-large risk premium, the student need only redefine the level for initial wealth to W* rather than W_0. The individual should then be forced to undertake the binomial gamble of being at either W_0 + A, or W_0 – B each with 50% probability. Since our initial wealth is W*, then this gamble is actuarially neutral since W*- (W_0-B) = (W_0+A) – W* with the left side being the possible amount to be lost, and the right side being the possible amount to be won.[5] The right side of equation (5) still measures the expected utility of this gamble, but W_0, being on the left side of (5), now becomes the certainty equivalent wealth level for this gamble. As indicated, W_0 allows the expected utility, given by (5), to equal the utility of the actuarially neutral gamble of starting at initial wealth W* and resulting at either W_0+A or W_0-B. The Markowitz risk premium measures the difference between the expected wealth (W* for this actuarial-neutral gamble) and the certainty equivalent wealth (W_0 in this example). As a result, (7) measures the Markowitz risk premium π associated with this gamble given that W* is the new initial wealth.

$$\pi = W^* - W_0 = (1/2)(W_0 + A) + (1/2)(W_0 - B) - W_0 = (A - B)/2 \quad (7)$$

[4] Note that (5) and (6) require this intersection.

[5] Equation (6) requires $2W^* = W_0 + A + W_0 - B$ which results in W* - (W_0–B) = (W_0+A) – W*.

The value π also has an additional interpretation that applies to the case when the initial wealth is W_0 rather than W*. If the individual starts at initial wealth W_0 (rather than W*), and undertakes the 50% - 50% gamble for either +A or –B, then given (7), π becomes the expected payoff for the gamble.[6] W* becomes the expected resulting wealth that is necessary to induce indifference between taking the gamble or not. The individual is indifferent between either staying at the initial wealth W_0 with certainty, or accepting the gamble with resulting wealth of $W^* > W_0$.

In the above analysis, π can therefore be interpreted as either (i) the expected value of the actuarially non-neutral gamble if W_0 is the initial wealth, or π can be interpreted as (ii) the Markowitz risk premium for the actuarially neutral gamble if W* is the initial wealth. It is shown below that both interpretations (i) and (ii) are useful for undergraduate exposition depending upon the context of the problem examined.

A Concavity Measure, Risk, and the Risk Premium

Given that our gamble is not normally distributed, but rather binomially distributed, and that we have not assumed a quadratic utility function for wealth, we know that the expected value and standard deviation are not the only relevant parameters for analysis of choice under uncertainty.[7] Nevertheless, the undergraduate curriculum generally assumes that these two parameters are sufficient for this analysis. For the 50% - 50% gamble, the standard deviation is given by (8).[8]

$$\sigma = (A+B)/2 \tag{8}$$

Given (7) and (8), the relation between π and σ is given by (9), which has a particular interpretation with respect to the concavity of the utility function.

$$\pi = \sigma(A - B)/(A + B) \tag{9}$$

As shown in Figure 1, the difference A – B is the amount that must be added to the winning stakes in order to induce indifference between the individual either taking the gamble or remaining at initial wealth W_0 with certainty. The sum of A + B is the spread between the gamble's outcomes. The undergraduate student should accept that the ratio (A-B)/(A+B) is a natural concavity measure in that the more concave the utility function is, the greater the amount that must be added to the winning stake as a portion of the total stakes. This ratio is always positive for a risk averse individual (by the definition above, A > B for risk aversion).

It is natural to compare this to Pratt and Arrow's in-the-small risk premium given by equation (3), where the ratio $-U''/U'$ is the concavity measure. The in-the-large risk premium presented by π, and the in-the-large concavity measure of (A-B)/(A+B), are clearly the more appropriate measures for the undergraduate curriculum. They are derived using simple algebra and geometry rather than the Taylor series expansion required for Pratt and Arrow's measure. For further insights into the values for A – B and π, and how they are related to Pratt and Arrow's in-the-small measure, a Taylor series expansion of these values is provided in the Appendix.

Numerical Illustrations with Revealed Preference

For a classroom illustration of our risk premium analysis, assume an individual's utility function is logarithmic where U = ln(W), and assume that the initial wealth is \$100. Also assume a 50% - 50% gamble with losing stakes of B as given in Table 1. We envision asking this individual for the corresponding levels for A that leave the individual indifferent about staying at the initial wealth level of \$100, or undertaking the gamble. Assuming the specification of the logarithmic utility function, the necessary values for A are also presented in Table 1. Given the values for A and B, then the resulting values for the risk premium π, the risk level measured by σ, and the concavity measure (A-B)(A+B), are also presented.

Table 1
Values for π Given U = ln(W), and W_0 = \$100

---

[6] Note that $E(W) = (1/2)(W_0+A) + (1/2)(W_0-B) = W_0 + (A-B)/2 = W_0 + \pi$ where (7) defines π.

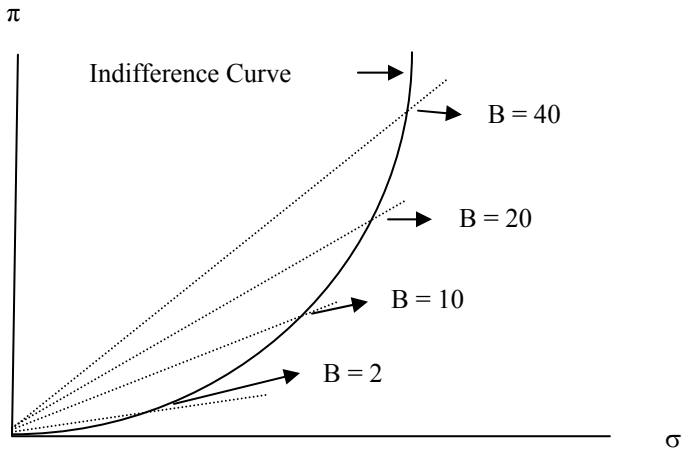
[7] Copeland and Weston (1983, Chapter 6) present one of many advanced text reviews of the requirements for mean and variance analysis. Either normality or a quadratic utility function is required for the standard deviation to be the appropriate risk measure.

[8] Note that $Var(W) = (1/2)\{W_0+A - (W_0 + (A-B)/2)\}^2 + (1/2)\{W_0-B - (W_0 + (A-B)/2)\}^2 = (1/4)(A+B)^2$.

B	A	π	σ	(A – B)/(A + B)
$ 2.00	$ 2.04	$ 0.02	$ 2.02	.0100
10.00	11.11	0.56	10.56	.0530
20.00	25.00	2.50	22.50	.1111
40.00	66.67	13.33	53.34	.2500

In Table 1, the initial wealth level stays at $100, but as the losing stake B increases, we note that the risk premium also increases, and more importantly for our analysis, the concavity measure also increases. For each point in Table 1, equation (5) is maintained at $U(W_0) = U(\$100)$. It is appropriate, therefore, to interpret each point as being along a constant utility indifference curve in the space of π and σ. In addition, the concavity measure $\pi/\sigma = (A-B)/(A+B)$ is the slope of the ray from the origin to the particular point along the indifference curve. As Table 1 shows, the increasing slope illustrates that the indifference curve is concave from above. This is illustrated by Figure 3.

Figure 3
Indifference Curve for $EU = U(W_0) = U(\$100)$
With Increasing Stakes B



It is important to note that a series of revealed preference questions can elicit data such as that presented by Table 1, but without the specification of the individual's initial wealth or utility function. This can be conducted as a classroom exercise by asking appropriate questions of students. One need only ask a selected student, "Given a 50% - 50% gamble where you could lose $2, how much must the winning stakes be in order to induce you to undertake the gamble?"[9] Once this answer is elicited (the value for A in Table 1), then the losing stakes can be varied, and for each elicited pair for A and B, equations (7), (8) and (9) can be used to complete the table. Knowledge of W_0 and the functional form for U are not necessary. The student's response implicitly incorporates this information. As an econometric exercise, students can use least squares to fit a particular functional form, such as the logarithmic, quadratic or exponential utility functions through the elicited data.

Risk Premiums with Futures Market Illustrations

One of the fundamental tasks required for explaining the functioning of forward and futures markets is to show that these markets are dominated by hedgers; that under conditions of homogeneous expectations, hedgers outbid

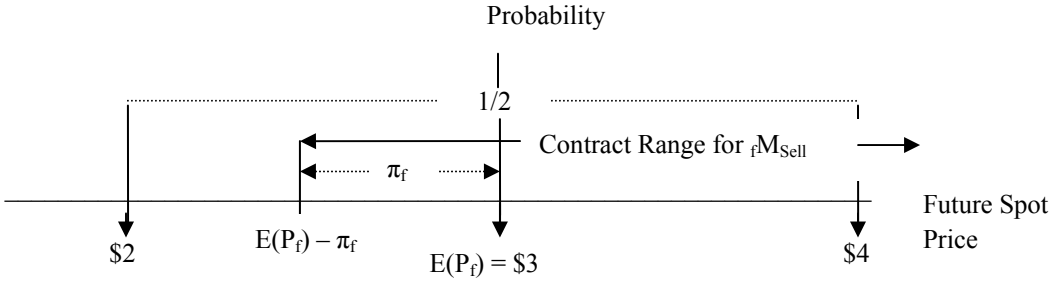
[9] It may be necessary to try more than one student in order to find one that is risk averse.

speculators. Speculators must expect to make profits that are sufficiently large to compensate for the risk. If a speculator is to sign a forward contract to purchase some good, then that contract price must be below what they expect the future spot price to be, and sufficiently below so that the expected profit is large enough to compensate for the risk. Likewise, if a speculator is to sign a forward contract to sell some good, then the contract price must be above the expected future spot price, and sufficiently above so that the expected profit is able to compensate for the risk. The minimum expected profit in each case is a Markowitz risk premium. Speculators absorb risk in return for expecting to make a profit, but hedgers expect to pay to eliminate risk.

Hedgers who have the good, and wish to sell it in the future, are willing to sign forward contracts for prices below what their expectations are for the future spot prices. They are willing to contract for a price below what they expect to get if they wait for the future spot market to develop. The difference between their expected future spot price and the contract price they are willing to accept is their Markowitz risk premium. Hedgers who want to purchase the good in the future are willing to sign forward contracts for prices greater than what they expect to receive if they wait for the future spot market to develop. The difference in price is also a Markowitz risk premium. The expected future spot price, acting together with the two risk premiums, establish the outer ranges for the forward contracts to buy and sell the good.

Forward and futures markets for simple agricultural commodities provide ready practical classroom illustrations for the use of the Markowitz risk premium. As an initial exposition of this topic, they are more easily understood than financial futures. For example, consider the winter wheat farmer with a crop to be harvested in three months. Allow the farmer's probability distribution for the future spot price to be binomial with price at future time 1, indicated by P_1, of either \$2 or \$4 with probability of ½ for each. This is illustrated by Figure 4. For this case, the farmer's expected price is $E(P_f) = \$3$, and $E(P_f) - \pi_f$ is the farmer's certainty equivalent price where π_f is the Markowitz risk premium.

Figure 4
Probability Distribution for Farmer's Future Spot Price



To establish this certainty equivalent price, we assume that the farmer's initial wealth (W_0) is \$100, and the farmer's utility of wealth function is $\ln(W_1)$. We also assume that the farmer has one unit of wheat for sale at either \$2 or \$4, the binomially random price with probability of ½ for each state. W_1 is therefore either \$104 or \$102 as dependent upon the sale price for the wheat. We therefore have $E(W_1) = \$103$. To find the Markowitz risk premium, we must first find the farmer's certainty equivalent wealth of $\$103 - \pi_f$. Since the utility of the certainty equivalent wealth must equal the expected utility of the binomial wealth distribution, then equation (10) implicitly defines π_f, which calculates to $\pi_f = \$.01$.

$$\ln(\$103 - \pi_f) = (1/2)\ln(\$104) + (1/2)\ln(\$102) \qquad (10)$$

Also allow $_fM_{Sell}$ to be the farmer's forward contract price to sell the crop with delivery at harvest. The risk averse farmer would sign any forward contract with price of $_fM_{Sell} > E(P) - \pi_f$ as illustrated by the "contract range" in Figure 4. The minimum contract price for the farmer is then \$2.99. This is so because at any price above the certainty equivalent, the utility of signing the contract, and therefore of receiving the certain forward price, exceeds the utility of being unhedged and facing the gamble.

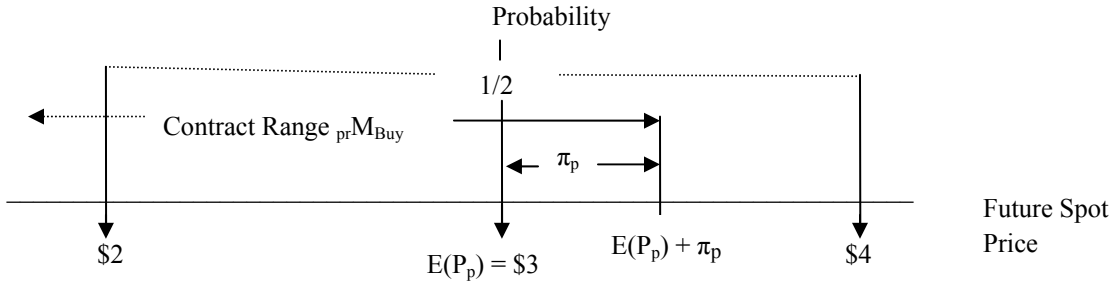
Consider the wheat processor who has the same probability distribution in mind for the future spot price as the wheat farmer (either \$2 of \$4 with probability of ½) where $E(P_p)$ is the processor's expected future spot price, and

$E(P_p) = E(P_f) = \$3$ as under conditions of homogeneous expectations. We allow π_p to be the Markowitz risk premium so that $E(P_p) + \pi_p$ is the certainty equivalent price for the processor. For simplicity, we assume that processing the wheat into a finished retail product is costless, and that the finished product will be sold for a certain price of \$5. The processor will process one unit of wheat so that the profit is $\$5 - P_1$. The processor's profit is therefore binomial at either \$5 - \$4 = \$1, or \$5 - \$2 = \$3 each at probability of ½. We also assume that the initial wealth of the processor is $W_0 = \$100$, so that future wealth W_1 is binomial at either \$101 or \$103 with probability of ½, and therefore $E(W_1) = \$102$. The processor's certainty equivalent wealth is therefore $\$102 - \pi_p$. Equation (11) implicitly defines π_p, which calculates to $\pi_p = \$.01$.

$$\ln(\$102 - \pi_p) = (1/2)\ln(\$101) + (1/2)\ln(\$103) \qquad (11)$$

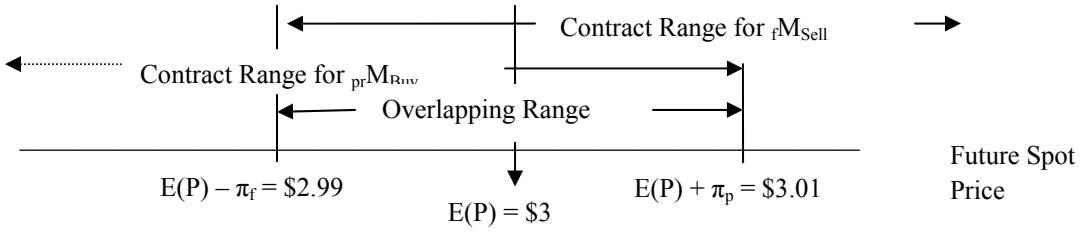
Allow ${}_{pr}M_{Buy}$ to be the processor's possible forward contract price for purchasing the unit of wheat at time 1. Being risk averse, the processor's contract range for acceptable forward contract prices to purchase the wheat is ${}_{pr}M_{Buy} < E(P_p) + \pi_p$. This indicates that the maximum acceptable contract price for the processor is \$3.01. Figure 5 illustrates the situation.

Figure 5
Probability Distribution for Processor's Future Spot Price



If the two contract ranges overlap, then a broker should be able to bring the processor and farmer together. They will overlap under conditions of homogeneous expectations ($E(P_f) = E(P_p)$) and risk aversion. Figure 6 illustrates the overlapping contract regions for the data analyzed above. As shown, under conditions of homogeneous expectations, brokers should be able to sign contracts with both processor and farmer for a maximum profit of $E(P_p) + \pi_p - E(P_f) - \pi_f = \$.02$.

Figure 6
Overlapping Contract Ranges Under Conditions of
Homogeneous Expectations



Unlike the farmers and processors, speculators must expect to make a profit from the pure purchase and sale of the product, not from growing and selling, or from purchasing, processing and selling. If $E(P_S) = E(P_f) = \$3$ is the speculator's expected price, then for the speculator to expect to make a profit from a forward contract to sell, she

must have the contract price ${}_{Spec}M_{Sell}$ within a range such that ${}_{Spec}M_{Sell} > E(P_S)$. In this case, the speculator expects to purchase at $E(P_S)$, and has a contract to sell at a profit of ${}_{Spec}M_{Sell} - E(P_S)$, this difference being the expected profit which must exceed or equal the Markowitz risk premium for the speculator to want to engage in the contract. Assume that the speculator also has a risk premium of $.01, and an expectation of $E(P_S) = \$3$. Figure 7 shows that contract range for the speculator to sell.

Under conditions of homogeneous expectations, however, the farmer being a hedger, should be willing to sign a contract to sell below this, at ${}_{f}M_{Sell} < {}_{Spec}M_{Sell}$. The farmer will outbid the speculator for contracts to sell. Figure 8 shows both the farmer's and speculator's contract ranges with homogeneous expectations. It illustrates that the farmer (hedger) will outbid the speculator for contracts to sell.

Figure 7
Contract Ranges For Speculator Under Conditions of
Homogeneous Expectations

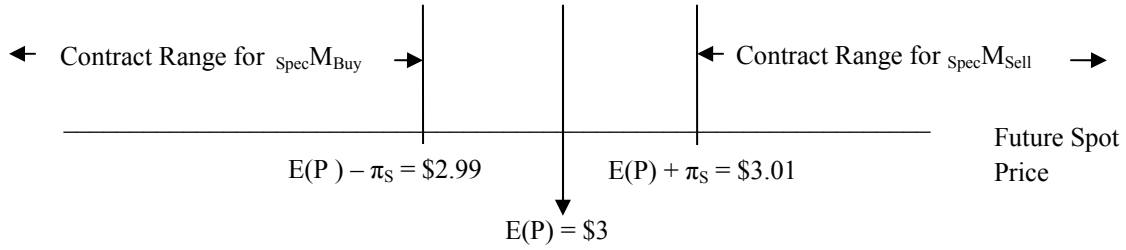
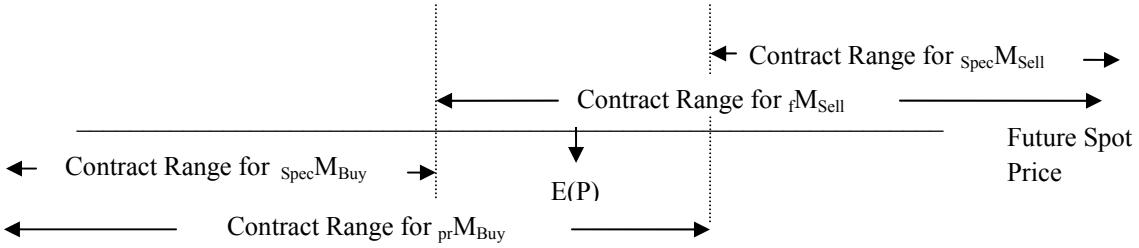


Figure 8
Contract Ranges For Speculator, Farmer and processor Under Conditions of
Homogeneous Expectations



Similarly, for the speculator to be willing to sign a contract to purchase the wheat, the contract price must be ${}_{Spec}M_{Buy} < E(P_S)$. In this case, the speculator expects to sell at a price above the forward contract price paid to purchase the wheat. Figure 7 illustrates the range for the speculator's contract for purchase. Being a hedger, the processor is willing to outbid the speculator and sign a contract to purchase at a price ${}_{Pr}M_{Buy} > {}_{Spec}M_{Buy}$, that is, sign a contract to purchase at a higher price. Figure 8 also shows the contract ranges for the speculator and processor under homogeneous expectations. It illustrates that the processor (hedger) outbids the speculator for contracts to purchase.

On both the purchasing side, and the selling side of the futures market, the speculator is outbid for contracts provided we have homogeneous expectations. Only under conditions of disagreement about the probability distribution for future spot prices would the speculator be able to play a role in this market.

To help reinforce these ideas, an appropriate project for the undergraduate is to allow the wheat processors and farmers to have the expectations and risk premiums as specified above, but to allow the speculator to disagree as to the probability distribution for the future price of wheat. If the speculator has a $.01 risk premium, what must the speculator's expected future spot price be to induce her to outbid the processor in signing a contract to purchase? What must the speculator's expected future spot price be to induce her to outbid the farmer in signing a contract to sell?

To answer the former question, since the processor is willing to pay \$3.01 on a forward contract to purchase, at this contract price the speculator must expect to sell at the future spot price of \$3.02 or above. To answer the latter question, since the farmer is willing to contract to sell at \$2.99, then to sign the contract to sell at this price, the speculator must expect to be able to buy at \$2.98 or below. These answers illustrate that only under conditions of disagreement between speculators and hedgers about the probability distribution for future spot prices would speculators be willing to sign contracts. Once the student realizes this, then the instructor can raise the possibility that either the speculator possesses superior inside knowledge, which is very unlikely in commodity markets, or that the speculator will make losses from her positions. The more likely case is the latter.

Summary of a Parsimonious Undergraduate Curriculum in Risk Theory

As based upon the above analysis, the following suggests an undergraduate curriculum for establishing a foundation in risk theory:

i) The utility of wealth, along with the notion of diminishing marginal utility, should be explored as the basis for choice under uncertainty. Possible mathematical forms for the utility function, such as $U = \ln(W)$ or $U = W^{1/2}$ where W is wealth, should also be explored.

ii) For 50% - 50% binomial gambles, the *fair utility axiom* (for indifference, the possible utility to be gained must equal the possible utility to be lost), must be developed. From this, the notion that risk aversion exists if A > B (as explained above) should be reviewed as sensible and workable, with the applicable concavity measure being the ratio A-B/A+B. It is then algebraically shown that this *fair utility axiom* requires that the expected utility of the gamble equals the utility of the initial wealth in order for indifference to occur, as shown by equation (5).

iii) The value for the Markowitz risk premium, π, should then be derived, and this should be related to equation (9) (π is directly related to the product of the concavity of wealth measure, and the standard deviation of the gamble). It should be fully explored that this risk premium represents the maximum the individual would pay to avoid this gamble, and that with π, we can establish the notion of the certainty equivalent wealth level.

iv) Revealed preference responses for A given B (as in Table 1 above) should be elicited from the undergraduates, and the measures for π and σ calculated. The notion of the indifference curve in the space of π and σ, and at a constant wealth level, should then be developed as based upon the revealed preference data. In class exercises that elicit student responses for A given various levels for B, and without specification of the utility function or the initial wealth level, but with purely subjective responses from the student, are all that are needed for development of π, and σ. Student comparisons of degrees of risk aversion (some may be risk lovers), are then appropriate to help reinforce these concepts.

v) To reinforce this risk analysis, applications to futures markets can be used to explain why hedgers, rather than speculators, dominate these markets when expectations are homogeneous.

Instructors can obtain a set of revealed preference exercises for classroom risk premium calculations by contacting the author by email. Futures markets exercises are also included.

Conclusion

Although expected utility and risk premia play a central role in finance theory, the mathematical sophistication of these subjects has resulted in their omission from the undergraduate curriculum. It is shown in the above analysis, however, that only simple algebra is necessary to derive both expected utility, and an alternative measure of in-the-large risk premium for 50% - 50% binomial gambles. This risk premium is shown to be a simple product of a geometric measure of the concavity of the utility function and the standard deviation of the gamble. Furthermore, this analysis is derived using only two very simple axioms of choice: (1) more wealth is preferred to less wealth, and (2) indifference requires that gambles must be fair utility wise.

Although these axioms do not penetrate the problems associated with either the independence axiom, or the comparability or ranking axioms, problems explored in advanced financial economic theory that are still being debated in the literature after 60 years, they are suitable for the undergraduate curriculum. Particular applications to futures markets help reinforce this analysis. This establishes a needed pedagogy in risk theory suitable for the undergraduate curriculum.

References

Arrow, K. J., *Essays in the Theory of Risk Bearing*, North-Holland, Amsterdam, 1971.

Copeland, T. and J. F. Weston, *Financial Theory and Corporate Policy*, Second Edition, Addison Wesley, Reading, Mass., 1983.

Friedman, M., and L. J. Savage, "The Utility Analysis of Choices Involving Risk," *The Journal of Political Economy*, August, 1948, 279-304.

Herstein, I. N., and J. Milnor, "An Axiomatic Approach to Expected Utility," *Econometrica*, April 1953, 291-297.

Machina, M., "Risk, Ambiguity, and the Rank-dependence Axiom," *American Economic Review*, March, 2009, 385-430.

Markowitz, H., *Portfolio Selection*, Yale University Press, New Haven, 1959.

Pratt, J. W., "Risk Aversion in the Small and in the Large," *Econometrica*, January-April, 1964, 122-136.

Von Neumann, J., and O. Morgenstern, *Theory of Games and Economic Behavior*. Princeton University Press, Princeton, N. J., 1947.

Appendix: A Taylor-Series Derivation of A – B

Equation (5) is repeated below. Equations (A1) and (A2) present Taylor series expansions of $U(W_0 + A)$ and $U(W_0 – B)$ about the point W_0. Equation (A3) therefore presents an approximation of $\{\cdot\}$.

$$U(W_0) = (1/2)U(W_0 + A) + (1/2)U(W_0 – B) = (1/2)\{\cdot\} \quad (5)$$

$$U(W_0 + A) \approx U(W_0) + U'A + (1/2)\, U''A^2 + \cdots \quad (A1)$$

$$U(W_0 - B) \approx U(W_0) - U'B + (1/2)\, U''B^2 + \cdots \quad (A2)$$

$$\{\cdot\} \approx 2U(W_0) + U'(A – B) + (1/2)U''(A^2 + B^2) \quad (A3)$$

Substitution of (A3) into (5) results in (A4), from which (A5) follows. Equation (A6) then follows from (9), which is also presented. Equations

$$0 \approx (1/2)\, U'(A – B) + (1/4)\, U''(A^2 + B^2) \quad (A4)$$

$$A – B \approx - (1/2)(U''/U')(A^2 + B^2) \quad (A5)$$

$$\pi = \sigma(A – B)/(A + B) \quad (9)$$

$$\pi \approx - (1/2)\sigma(U''/U')(A^2 + B^2)/(A + B) \quad (A6)$$

The Relationship between the Promised and Realized Yields to Maturity Revisited

***Hassan Shirvani*[1], *Barry Wilbratte*[2]**

ABSTRACT

This paper reaffirms the long-held view that the promised yield to maturity of a coupon bond can be realized only under certain restrictive conditions. Specifically, the realized yield equals the promised yield only if the spot rate and yield curves are flat and remain unchanged throughout the term of the bond, a condition which rarely if ever holds. In addition, we explain that, regardless of the reinvestment rates, the promised yield on a coupon bond will be realized, provided that the bond is held not to its maturity but to its duration. (Key words: promised yield to maturity; realized yield; spot rate curve; yield curve; duration)

Introduction

In a recent pair of companion papers in this journal, Cebula and Yang (2008) and Forbes, Hatem and Paul (2008) take issue with the long-standing proposition that the promised yield to maturity of a coupon bond held to maturity is realized only if the bond coupons are reinvested at the same rate as the promised yield (Fabozzi and Modigliani, 2002; Mayo, 2008). The papers assert, using somewhat different approaches to justify this assertion, that the promised yield to maturity of a coupon bond is always earned regardless of whether or not those coupons are reinvested. To show this, the papers draw a distinction between the concepts of the yield to maturity (YTM) and the realized compounding yield (RCY) for a coupon bond. The authors argue that the YTM, defined as the discount rate that equates the present values of the cost and benefits of the bond, is earned whether or not the bond coupons are reinvested. In contrast, the RCY is earned only under specific assumptions regarding how those coupons are reinvested.

This paper argues that, contrary to the assertions of the aforementioned articles, the YTM of a coupon bond, which is based on the unrealistic assumption of a flat and unchanging spot rate curve (the yield curve for zero coupon bonds of various maturities, Lyuu, 2002) and, correspondingly, a flat yield curve, is a fictitious mathematical construct, never to be actually realized until and unless the bond coupons are all reinvested at the same unchanging current interest rate. In other words, the YTM is a forward-looking approximation to the true rate of return on a coupon bond, which is a backward-looking concept measured by the RCY. Since the assumption of a flat and constant yield curve runs counter to most and possibly all real life experience, it is clear that the promised YTM will virtually never be realized, and, is therefore devoid of much economic significance. In fact, only for zero coupon bonds, which have no reinvestment rate problem, is the YTM a real and realizable measure over the life of the bond.

Furthermore, should the assertion that the YTM has an existence independent of the reinvestment rate of the coupons be true, it immediately follows that all otherwise identical bonds with different coupon rates must offer the same YTM. For example, suppose there are two equally risky 10-year bonds on the market but with very different coupon rates. The entities which issued these bonds would have a common borrowing rate for a given class of securities and thus, by the above reasoning, these bonds should offer the same YTM. Empirically, however, it is often the case that otherwise identical bonds with different coupon rates differ in YTM (Caks, 1977). Of course, differential tax rates on coupon incomes and capital gains also contribute to the differential yields (Robickek and

[1] Professor of Economics and Finance, University of St. Thomas, Houston, TX 77006, email Shirvani@stthom.edu
[2] Professor of Economics and Finance, University of St. Thomas, Houston, TX 77006, email wilbratt@stthom.edu

Niebuhr, 1970). But it is also known that investors, in an attempt to minimize the adverse effects of falling reinvestment rates on their realized yields, tend to favor, and hence to accept lower yields, on bonds with lower coupon rates in an environment of falling interest rates (Buse, 1970). Thus, the YTM and the reinvestment rate cannot be decoupled without ignoring the realities of an efficient bond market. In addition, the fact that otherwise identical bonds with different coupon rates can offer different yields renders the traditional yield curve of limited use in determining the fundamental value of a bond, which is the present value of the individual coupons and par of a bond, each discounted at the zero-coupon rate corresponding to the timing of each coupon bond payment.

The fundamental values of two bonds with different coupon rates will reflect different yields. For example, suppose the yield curve implies a yield to maturity of 10 percent for a 30-year Treasury bond with a 5 percent coupon rate. This yield cannot then be used to determine the value of a similar bond with a different coupon rate such as 7 percent. This is why we need the concept of the spot rate curve mentioned earlier. This curve, which is the yield curve of zero-coupon bonds of identical risk and different maturities, can be used to determine the fundamental value of any coupon bond as a package of zero coupon bonds, where the individual coupons and the par are treated as a sequence of zeros. Once the spot rate curve has been used to determine the value of a coupon bond, this value can then be used to compute the YTM of this bond. It should be noted that, although the yield and spot rate curves are not identical, they tend to have similar shapes. In particular, a flat spot rate curve necessarily indicates a flat yield curve.

On the other hand, as is well-known, the promised yield of a coupon bond will always be realized regardless of whether the coupon payments are reinvested or not, provided that the bond is held not to its maturity but to its duration (Shirvani and Wilbratte, 2002 and 2005). This result follows from the fact that any coupon bond with duration D is equivalent to a zero-coupon bond with maturity D (and the same price). To obtain this result, we note that a coupon bond can be immunized from reinvestment rate risk by setting its duration equal to a target investment horizon. Under these conditions, the immunized bond will yield the same lump at point D (FV_D) regardless of changes in interest rates because variation in the end price of the bond will offset reinvestment rate effects. This means that the coupon bond is equivalent to a zero-coupon bond with maturity D and par value FV_D. This equivalence means that if such a coupon bond is held for D years, this action produces the same yield, to a very close approximation, as holding the corresponding zero-coupon bond until its maturity D. As the promised yield to maturity of any zero-coupon bond held to its maturity will always be realized, so will the promised yield on the corresponding coupon bond. However, our disagreement with the papers is the much stronger assertion that the promised yield to maturity of a coupon bond will always be realized, even if that bond is held to maturity.

The remainder of this paper is organized as follows. Section II discusses the nature of the concept of the yield to maturity at some length to identify the conditions necessary for its realization. This section also briefly examines some of the arguments advanced in the papers cited above. Section III concludes.

The yield to maturity

The concept of the yield to maturity (YTM), a purported measure of the average annual rate of return of a bond held to maturity, is defined as the discount rate that equates the present values of the cost and benefits of a bond, as shown below:

$$P = \sum_{1}^{N} \frac{Ct}{(1+YTM)^{t}} + \frac{F}{(1+YTM)^{N}} \qquad (1)$$

where P = the current bond price, C_t = the coupon payment in year t, F = the face value of the bond, and N = the bond maturity. Since the above equation is based on the unrealistic assumption that the yield curve is flat and remains both flat and constant for the term of the bond, it is clear that the yield to maturity concept is a fictitious measure of the total bond return, suitable only for ordinal rankings of bonds with similar characteristics. Put differently, it is clear from the above equation that in general there is no guarantee that the promised yield to maturity will ever be actually earned. To see this, we convert the above equation from its present value version into a future value version by simply multiplying both sides of the above equation by $(1 + YTM)^N$ to reach the expression

$$P(1+YTM)^{N} = \sum_{1}^{N} Ct(1+YMT)^{N-t} + F \qquad (2)$$

Written in this alternative but equivalent form, the equation now defines the yield to maturity as the rate of return that equates the future (reinvested) values of the cost and benefits of the bond. In particular, the equation shows that for YTM to be realized, so that YTM = RCY, it is necessary for all future coupons to be reinvested at the same rate of return of YTM, that is, the yield curve must remain flat and constant at the initial yield to maturity of YTM over the entire life of the bond. If, instead, future interest rates change, then YTM will not be realized as promised, resulting in a divergence between YTM and RCY.

Clearly, the authors of the papers cited in our introduction recognize that the actual realized yield will in general differ from the promised yield to maturity, the variation from that yield depending upon the reinvestment rate. The reasoning of their papers, however, submerges the fact that the yield to maturity has no existence independent of the future reinvestment rates. In other words, while the concept of the realized yield is always real, the concept of the yield to maturity is in general fictitious, becoming a reality only in the rare and perhaps nonexistent case in which the current interest rate persists until the bond matures.

The meaning of the previous paragraph can perhaps be best clarified by reference to a particular example in the second of the above papers. For a 5-year, 5 percent coupon bond quoted at $1,000, the paper derives a yield to maturity of 5 percent. The paper then proceeds to decompose the bond price into a set of present values of the future benefits of the bond. For example, the present value of the first coupon payment on the bond ($50) discounted at 5 percent amounts to $47.62, the present value of the second coupon amounts to $45.35, and so on. Based on these results, the paper asserts that the first component ($47.62) of the $1,000 bond price has earned 5 percent return for one year, the second component ($45.35) has earned 5 percent for two years, and so on. Now, since every dollar of the bond price has earned a return of 5 percent, the paper concludes, it follows that the total rate of return for the bond as a whole over its life of 5 years is also 5 percent.

The problem with this argument is that the different components of the price of the bond have not earned a 5 percent return over the life of the bond, a condition which is necessary for concluding that the rate of return for the entire bond price is also 5 percent. Specifically, the first $47.62 has earned 5 percent for only one year, the second $45.35 for only two years, and so on. For any of these components to earn exactly 5 percent return over the entire life of the bond, it is necessary that each of them be invested for 5 years. Furthermore, for each of these components to earn exactly 5 percent over the life of the bond, the paper assumes that the reinvestment rate remains constant at 5 percent for the next five years, which in turn occurs only if the spot rate and yield curves are flat and unchanging. Thus, as asserted earlier, without knowledge of the reinvestment rates for the next five years, it is simply impossible to determine the total rate of return for the bond held to its maturity.

To further clarify the preceding point, consider also the bond in question under a different spot rate and yield curves. Specifically, let us now assume that the spot rate curve slopes upward, with long term interest rates with maturities one, two, three, four, and five years to be, respectively, 5, 6, 7, 8, and 9 percents. Under these conditions, the first coupon payment of $50 is now discounted once at 5 percent to yield a value of $47.62, the second payment is discounted twice at 6 percent to yield a value of $44.50, and so on. Summing up these discounted values, we obtain a fundamental value of $852.11 and a corresponding YTM of 8.78 percent for the bond. In addition, and following the earlier argument, we can say that the first cash flow of the bond has earned a return of 5 percent for one year, the second has earned a return of 6 percent for two years, and so on. Clearly, none of these partial returns, which again cannot be used to determine the total return of these separate components over the entire life of the bond, bears any relationship to the promised yield to maturity of 8.78 percent for the bond as a whole. What is more, under our assumption of an upward sloping yield curve, there are no circumstances under which the promised yield to maturity of 8.78 percent can ever be realized for the bond in question, simply because we have ruled out the flatness of the yield curve by assumption.

Conclusion

This paper defends the long-held proposition that the promised yield to maturity of a coupon bond held to its maturity is in general unrealizable, unless the yield curve is assumed to be both flat and unchanging, a condition which rarely if ever holds in financial markets. In particular, the present paper questions the recent assertion by two articles recently published in this journal that the promised yield to maturity has an existence independent of the shape and position of the yield curve. At the same time, this paper shows that while the promised yield to maturity of a coupon bond is generally unrealizable if the bond is held to its maturity, it is attainable if the bond is held to its duration.

References

Buse, A.1970. "Expectations, Prices, Coupons and Yields," *Journal of Finance,* 809-818.

Caks, J.1977. "The Coupon Effect of Yield to Maturity," *Journal of Finance,* 103-115.

Cebula, Richard J., and Bill Z Yang. 2008. "Yield to Maturity Is Always Received As Promised." *Journal of Economics and Finance Education* 7, 43-47.

Fabozzi, Frank, J. and Franco Modigliani. 2002. *Capital markets – Institutions and Instruments*, 3rd edition, Prentice Hall.

Forbes, Shawn M, John J. Hatem, and Chris Paul. 2008. "Yield-to-Maturity and the Reinvestment of Coupon Payments." *Journal of Economics and Finance Education* 7, 48-50.

Lyuu, Yuh-Dauh. 2008. *Financial Engineering and Computation Principles*. Cambridge University Press.

Mayo, Herbert. 2008. *Investments: An Introduction*, 9th Edition, United States.

Robickek, A. and W. Niebuhr. 1970. "Tax-Induced Bias in Reported Treasury Yields," *Journal of Finance,* 1081-1090.

Shirvani, Hassan and Barry Wilbratte. 2002. "Two Pedagogical Simplifications of the Concept of Duration." *Journal of Economics and Finance Education* 7, 18-23

_____. 2005. "Duration and Bond Price Volatility: Some Further Results." *Journal of Economics and Finance Education*, 1-6.

Yield to Maturity Is Always Received as Promised: A Reply

Richard Cebula[1] and Bill Z. Yang[2]

ABSTRACT

This note attempts to further spell out why it is a myth that YTM is viewed as only a promised but not really earned interest rate. It addresses some misconceptions in Shirnani and Wilbratte (2009, this issue) on what, between YTM and RCY, is a true rate of return of a coupon bond, why YTM is NOT just a "fictitious mathematical construct", and why YTM has nothing to do with yield curve.

Introduction

There has been a long-held myth in financial economics that the yield to maturity (YTM, hereafter) of a coupon bond is only promised and may not be actually earned unless coupon payments are *reinvested* at the same interest rate as the initial YTM (e.g., see Reilly and Brown, 1997, pp. 530-531, and Strong, 2004, p.70, among others). Recently, we have formally demonstrated why it is indeed a myth (Cebula and Yang, 2008). In that paper, we rigorously proved a proposition that the (initial) YTM when purchasing a coupon bond measures the interest rate actually earned from holding the bond until maturity (assuming no defaults); furthermore, the realized compounding yield (RCY hereafter) is in fact the YTM of a portfolio that holds the bond and reinvests some or all coupon payments as they are received.

Attempting to defend *the* long-held myth, Shirvani and Wilbratte (2009, this issue) criticize that the (initial) YTM "is based on the unrealistic assumption of a flat and unchanging yield curve", and hence it is only "a fictitious mathematical construct, never to be actually realized until and unless the bond coupons are all reinvested at the same unchanging current interest rate." Moreover, they propose that "the true rate of return on a coupon bond … is …the RCY." We understand that we should not expect everyone to accept a correct but unconventional point, in particular, since the conventional myth has been so deeply rooted in many textbooks as well as in professionals' hearts. However, we find that Shirvani and Wilbratte have not only missed the major point we have made in our paper, but also have made some incorrect assertions in their comment. In this reply, we endeavor to address these issues.

Is YTM Fictitious or Real? A Counter Example

The key difference between the long-held myth and our proposition lies in what is the correct measure of the interest rate from holding a coupon bond until maturity. We believe that it is the YTM, defined as the solution for the following equation:

$$P_0 = \sum_{t=1}^{N} \frac{C_t}{(1+YTM)^t} + \frac{F}{(1+YTM)^N} \qquad (1)$$

where P_0 = the purchase price of the coupon bond, C_t = coupon payment per period, F = face/par value of the bond, and N = term to maturity. Note that a bond contract, in theory and in practice, is completely

[1] Richard J. Cebula, Shirley and Philip Solomons Eminent Scholar Chair and Professor of Economics, Armstrong Atlantic State University, Savannah, GA 31419, Richard.cebula@armstrong.edu

[2] Bill Z. Yang, Associate Professor of Economics, Georgia Southern University, Statesboro, GA 30460-8151, billyang@georgiasouthern.edu

characterized by parameters {F, C_t, N}, and that its purchase price is determined in the bond market by demand and supply. By definition, the YTM is internally determined by the foursome parameters {F, C_t, N; P_0} at the time of purchase. Hence, as long as there is no default, interest is paid exactly on time, and the bond is held until maturity, the YTM is not only a promised, but also actually earned (annual) rate of return by the bond holder until maturity. In this definition, nothing is assumed about how coupon payments are used – reinvested or spent, and nothing is assumed how the *current* market interest rate will vary over time after purchase. What is more, nothing is claimed on how much ending worth in year N for the bondholder will be built up, which relies on how coupon payments are managed – an additional investment by the bondholder.

In contrast, Shervani and Wilbratte (2009) suggest that "the true rate of return on a coupon bond … is a backward-looking concept measured by the RCY". Here, RCY stands for *realized compounding yield*, defined as the solution from the following equation:

$$P_0 (1 + RCY)^N = V_N \quad (2)$$

given P_0 = the initial investment (i.e., the purchase price of the bond in this case), and V_N = the ending-worth value at the end of year N. By nature, RCY measures the annual rate of return for an investment account, rather than for a single investment. In this definition, it is implicitly assumed that there is no leakage from, or injection into, the account over the investment horizon. For example, if the investment is coupon bonds and when coupon payments are received, they are assumed to be reinvested at a *current* interest rate that the investor can obtain. Therefore, V_N and hence, RCY, will be determined by the interest rate of *reinvestment*. For convenience of presentation, technically, if the coupon payments are spent rather than reinvested, we interpret it as reinvesting at a rate of negative 100% (i.e., all coupon payments are completely lost in the reinvestment). In addition, we exclude the possibility that additional funds other than the coupon payments can be added to the account – a case of a *de facto* Ponzi scheme.

Since our first point in this note is to refute that "the true rate of return on a coupon bond … is a backward-looking concept measured by the RCY" (Shervani and Wilbratte, 2009), we only need to give a counter example. In fact, we have already provided such an example in our paper (2008). Unfortunately, it was apparently overlooked. Hence, we cite it here.

Consider two investors, A and B. Each has bought the same coupon bond with F = \$1,000 and C_t = \$50 at a price of par value, i.e., P_0 = \$1,000. Besides, each has held the bond until maturity, e.g., N = 10 years. Clearly, both investors have put \$1,000 down and earned \$50 from it every year for 10 years. Clearly, YTM = C_t/F = 5% for both investors.

Moreover, investor A has somehow managed to reinvest all couple payments at an interest rate of 5% whenever received, whereas investor B has relied on the coupon payments to pay bills. As a result, investor A is able to have gathered an amount of $\$1{,}000(1.05)^{10}$ in her investment account by the end of year 10, while investor B can only have redeemed the principal of \$1,000. Hence, RCY = 5% for investor A, and it is zero for investor B.

So, what is the (annual) rate of return actually earned from holding the bond for each investor in this example? We would tell both investors that "you both have earned an interest rate of 5% from holding the bond, even though you have managed coupon payments differently." A possible question from investor A could be "if we both have earned the same interest rate, how come I have much more ending wealth accumulated than he does?" We would answer that it is because investor A has saved more (by reinvesting the coupon payments) than investor B does, but not because *the* bond bought by investor A yields a higher return than the one held by investor B.

Now, if Professors Shervani and Wilbratte told investor B that "the so-called 5% interest rate as promised by YTM is only a fictitious mathematical construct. In fact, you have earned *nothing* from holding this bond, because you have spent all coupon payments", investor B may question "What about the \$50 dollars I have received every year for 10 years from my \$1,000 investment? Isn't it real, or is it just fictitious?" We don't know how to answer this question if we were asked, but we would never tell investor B that he has earned nothing, especially in view of Cebula and Yang (2008).[3]

[3] One of us must confess that he had taught *the* myth in his classes a couple of times until someday a student asked a similar question to one by investor B. Pretty often, we instructors could benefit a lot from "seemingly stupid" question that challenges a long-held convention. If we cannot answer them logically, it must be because either we are ignorant or something may be wrong in the long-held convention.

Obviously, there is something of a problem with using RCY as the measure of rate of return from holding a coupon bond. So, what is wrong? In our paper (2008), we have formally shown that the RCY does not measure the rate of return from holding a coupon bond *per se*; rather, it actually measures the YTM of a portfolio with two investments: holding the coupon bond until maturity *plus* investing all coupon payments when received in a time deposit compounded annually. Of course, the value built up at the end of year N will depend on how coupon payments are managed – reinvested or spent, and what interest rates are when coupon payments are reinvested. But the (initial) YTM is the same from holding the initial bond itself, no matter how coupon payments are managed after being dispersed to the bondholder. Intuitively, the RCY is a weighted average of the initial YTM and the interest rate of reinvestment; hence, not surprisingly, RCY = YTM *only if* the reinvestment interest rate equals to the initial YTM. The formal proof for this proposition can be found in our paper (2008). We must emphasize that the proposition we made in (Cebula and Yang, 2008) was not an "assertion", since we have rigorously proven it rather than simply claiming it without a proof.

YTM Has Nothing to Do with Yield Curve

In addition to the different perspectives on what, between YTM and RCY, correctly measures the interest rate from holding a coupon bond, there is another very basic misconception in Shervani and Wilbratte (2009), namely, that "the YTM of a coupon bond, which is based on the unrealistic assumption of a flat and unchanging yield curve, is a fictitious mathematical construct". In the previous section, we have employed a counter example to refute the assertion that "YTM is only a fictitious mathematical construct". Now, we clarify why the YTM has nothing to do with any assumptions on the *current* yield curve.

Recall that for a bondholder, the (initial) YTM at the time of purchase is defined in equation (1) and completely determined by parameters $\{F, C_t, N; P_0\}$. Once the bond is actually purchased, the resulting YTM is fixed *for the bondholder* until maturity, assuming that the bond is held until maturity without default. Indeed, after the bond is purchased, the *current* market YTM will continue to change over time along with the market value of the bond purchased. Also, the term "to maturity" of the bond purchased will get shorter and shorter as time goes on. However, the (initial) YTM has nothing to do with the *current* market interest rate *after* purchase, because it has been locked by $\{F, C_t, N; P_0\}$. This reasoning is the same as the fixed-term mortgage rate and the CD rate. The *current* market interest rates (of the mortgage and the CD) may change continually. Once a contract is signed, however, the rate becomes fixed for the persons who have signed the contract. We should not get confused between the *current* market interest rate, which is in change all the time, and the interest rate one actually receives/pays, which becomes fixed after signing the contract. A confusion between the two kinds of interest rates before and after a contract is signed is a very basic misconception.

Therefore, we don't need any additional assumptions on how the current market yield curve behaves when we define and calculate YTM. Likewise, we have not assumed whether the yield curve is flat and unchanging when we proved the proposition on how RCY is related to YTM in our paper (2008).

Concluding Remarks: More on YTM

YTM is an internally determined measure of (total) rate of return (IRR) of fixed-income investment instruments, by their cash in/out flows until maturity. By nature, it serves as a common yardstick when "an apple is compared to an orange".

One relevant question on bond investment is how to use coupon bonds to accumulate an ending-worth value of $P_0(1 + YTM)^N$ at time N, given an initial investment P_0 and the initial YTM of the coupon bond. Of course, it is not automatic to simply hold the bond; rather, it requires the investor or her account/portfolio managers to reinvest coupon payments when received. A sufficient condition is to invest all coupon payments in CDs or discount bonds at the initial YTM, compounded annually (Fabozzi and Modigliani, 2002, p. 364). But the reinvestment strategy has nothing to do with the initial YTM. Unfortunately, this real-world question and proposed solution for it have been misinterpreted as that the initial YTM of the coupon bond is only a promised "fictitious mathematical construct" unless the ending-

worth value of $P_0(1 + YTM)^N$ can be accumulated. This misinterpretation is a problem for financial economics professionals, because it makes people confused between two different issues: What is the annual rate of return from holding a coupon bond until maturity? (*The* answer: It is the YTM), and how to accumulate an ending-worth value of $P_0(1 + YTM)^N$ by using coupon bonds with an initial YTM? (*An* answer: If one can reinvest coupon payments when received at the initial YTM).

Our paper (2008) has formally shown that the RCY is no other than the YTM of a portfolio. With a counter example this note has further refuted that the initial YTM of a coupon bond is NOT a "fictitious mathematical construct". We must acknowledge that we have not formally addressed what the initial YTM (of a coupon bond) actually measures. This issue is beyond the scope of this short reply. Nevertheless, our attention is drawn to another "hypothesis" in financial economics. Although many authors and researchers claim that the YTM "is often viewed as a measure of the average rate for return that will be earned on a bond if it is bought now and held until maturity" (Bodie, *et al*, 2002, p. 426), this statement remains an "assertion." To our knowledge, to date, it has not actually been formally proven in the published literature.[4]

References

Bodie, Zvi, Alex Kane, and Alan J. Marcus, 2002. *Investments*, 5th edition, McGraw-Hill Irwin

Cebula, Richard J., and Bill Z. Yang. "Yield to Maturity Is Always Received As Promised." *Journal of Economics and Finance Education* 7, Summer 2008, 43-47.

Cebula, Richard, Xiezhang Li, X. Henry Wang and Bill Z. Yang, 2009. "Yield to Maturity and Total Rate of Return: A Theoretic Note," Unpublished Manuscript (available upon request)

Fabozzi, Frank, J. and Franco Modigliani, 2002. *Capital Markets – Institutions and Instruments*, 3rd edition, Prentice Hall.

Reilly, Frank K. and Keith C. Brown, 1997. *Investment Analysis and Portfolio Management*, 5th Edition, Dryden Publisher

Shirvani, Hassan and Barry Wilbratte. "The Relationship between the Promised and Realized Yield to Maturities Revisited." *Journal of Economics and Finance Education* 8, Summer 2009, pp-pp

Strong, Robert A., 2004. *Practical Investment Management*, 3rd edition, Thomas Southwestern

[4] A formal proof for how YTM and rate of return are related can be found in Cebula, *et al* (2009, available upon request.)

Yield-to-Maturity and the Reinvestment of Coupon Payments: Reply

Shawn M. Forbes, John J. Hatem and Chris Paul[1]

ABSTRACT

Our original note addressed a common misconception that to earn the yield-to-maturity (YTM) on a coupon bond an investor must reinvest the coupon payments. Shirvani and Wilbratte (2009) take issue with our presentation and results. We will demonstrate that their arguments entirety rest on the proposition that the YTM must equal the realized compounded yield (RCY). This is a construct that explicitly assumes coupon reinvestment. We made no claim in our original presentation with regard to their proposition, because it is not required to calculate the YTM. Furthermore, we will discuss their claims with regard to the "economic significance" of the yield to maturity measure.

Introduction

In a recent issue of this journal (Forbes, Hatem and Paul, 2008), we presented a simple demonstration that yield-to-maturity is a discount rate and that its calculation requires no assumption of coupon payment reinvestment. Our purpose was to correct a recurring error found in many contemporary investment texts in spite of Renshaw (1957) addressing the misconception for capital budgeting decisions over 50 years ago. In their comment Shirvani and Wilbratte (2009) provide two additional examples (Fabozzi and Modigilianai, 2002; and Mayo, 2008) of the coupon-reinvestment assumption. Their comment makes several arguments against our demonstration and additional claims concerning the usefulness of YTM; as will be shown below none of their arguments or claims are relevant to our original point.

Yield-to-Maturity is a Discount Rate

The standard definition of YTM describes the method of calculation. It states that, "The yield to maturity is the single discount rate that, when applied to all future interest and principal payments, produces a present value equal to the purchase price of the security." Note that the definition explicitly states that YTM is a **discount rate** and is used to equate future cash flows with the bond price or **present value**. No assumption or condition for the reinvestment of coupon payments is made or required. The calculation of YTM is definitional and neither controversial nor ambiguous. If all promised payments are received the bond purchaser will earn the YTM.

Yield-to-Maturity is Not a Compound Rate

By rearranging the present value equation into their equation (2), a calculation of future value, Shirvani and Wilbratte (2009) analysis imposes additional requirements unrelated to, and unnecessary for, the calculation of the YTM. They state that the equation is an "alternative but equivalent form." However, the forms are not equivalent from a financial perspective. Specially, the future value form imposes the additional requirement that the YTM must equal the RCY and the calculation of RCY requires coupon reinvestment. By imposing these additional conditions they are no longer discussing YTM as correctly defined and calculated. Consequently their conclusion that, "the promised yield to maturity of a coupon bond … cannot be meaningfully separated from the rates at which the bond coupons are reinvested," is simply incorrect. Their additional claim that, "while the promised yield to maturity of a coupon bond is

[1] Shawn M. Forbes, Professor of Finance, Department of Accounting and Finance, Wofford College, Spartanburg, South Carolina 29303-3663, John J. Hatem, Associate Professor of Finance, and Chris Paul, Professor of Finance, Department of Finance and Quantitative Analysis, Georgia Southern University, Statesboro, Georgia 30460-0001

generally unrealizable if the bond is held to maturity, it is attained if the bond is held to duration," is based on the reinvestment of coupon and the sensitivity of bond prices to changes in interest rates, both of which are unrelated to the calculation of YTM.

A bond is a debt instrument and has the same characteristics as an interest-only mortgage with a principal balloon payment. The coupon and the mortgage interest payments are calculated in the same manner. However, nowhere in the literature is it claimed that the mortgagor must reinvest the interest payments at the original mortgage rate until the principal is repaid at the time of maturity to earn the calculated mortgage rate.

The Economic Importance of Discount Rates

In addition to their technical arguments the authors also claim that because the YTM will seldom equal the RCY; it a "fictitious mathematical construct," and "is devoid of much economic significance." Rather than being fictitious the YTM specifies the relationship between a transaction price and the timing and magnitude of future real cash flows. Discount rates are used in all types of financial instruments with both lenders and borrowers relying on them for specifying loan terms. That is, the present value or loan amount and the timing and magnitude of payments. Additionally, these discount or interest rates for various types of credit are necessary for the efficient allocation of scarce financial capital in both the firm and economy. This is a function that the backward-looking RCY cannot perform. Thus, we argue that discount rates are both real and of economic significance.

Conclusion

We reaffirmed that the calculation of the yield-to-maturity does not require the coupon payment reinvestment assumption, and that the yield-to-maturity will be earned if all payments are made as promised. The critique offered by Shirvani and Wilbratte was shown to rest upon the equivalency the present value and future value calculations. While mathematically equivalent, the two equations are not financially equivalent, as the future value calculation imposes the additional conditions of coupon reinvestment and equality of the discount and compound rate. Thus, their conclusions are irrelevant to the correct calculation of YTM.

As the calculated discount rate of future cash flows the YTM is no different than other calculated discount rates used in every type of loan agreement. As such, the YTM and other calculated rates are indispensable in the operation of credit markets. Backward-looking compound rates cannot perform this function.

References

Fabozzi, Frank J. and Franco Modigliani, *Capital Markets – Institution and Instruments*, 3rd Edition, Prentice Hall.

Forbes, Shawn M., John J. Hatem, and Chris Paul. "*Yield-to-Maturity and the Reinvestment of Coupon Payments*." Journal of Finance and Economics Education, Volume 7, Number 1, Summer, 48-51.

Johnston, Ken, Shawn Forbes, and John Hatem, 2002. "*Reinvestment Rate Assumptions in Capital Budgeting: A Note*," Journal of Economics and Finance Education, Volume 1, Number 2, Winter, 28-29.

Mayo, Herbert, 2008, *Investment: an Introduction*, 9th Edition, United States.

Renshaw, Ed. 1957. "*A Note on the Arithmetic of Capital Budgeting Decisions*," Journal of Business, Volume 30, Number 3, July, 193-201.

Shirvani, Hassan and Bary Wilbratte, "*The Relationship between the Promised and Realized Yield to Maturities Revisited*." In this issue.