# A Primer and Assessment of Social Security Reform in Mexico

**MARCO A. ESPINOSA-VEGA AND TAPEN SINHA**

*Espinosa-Vega is a senior economist in the Atlanta Fed's research department. Sinha is Seguros Comercial America Chair Professor of Risk Management and Insurance at the Instituto Tecnológico Autónomo de México.* *They thank Asok Chaudhuri, Frank King, Dipendra Sinha, Selahattin Imrohoroglu, Steve Russell, and Steve Smith for insightful comments.*

WHILE A NUMBER OF THEORETICAL ECONOMISTS HAVE ACCEPTED THE NOTION THAT MOVING FROM A PAY-AS-YOU-GO TO A FULLY FUNDED SOCIAL SECURITY SYSTEM WOULD IMPROVE A COUNTRY'S WELL-BEING, THERE IS FAR FROM UNIVERSAL AGREEMENT ON THIS POLICY PRESCRIPTION.[1] FOR EXAMPLE, WHILE KOTLIKOFF (1996) CALLS FOR A move to a fully funded pension system in the United States, Diamond (1998) presents a number of caveats for such a move.[2] Moreover, the major thrust of the World Bank (1994) that advocates moves away from a pay-as-you-go system has been severely criticized by Orszag and Stiglitz (1999) from within the World Bank itself. Perhaps these discrepancies explain why, to date, only a few economies have switched from a pay-as-you-go to a fully funded system.[3] Recently, however, the economic projections of a number of countries with pay-as-you-go systems have shown significant future actuarial imbalances. As a consequence, several of these countries are either contemplating or are engaged in a significant redesign of their pay-as-you-go systems.

While in the United States the debate about switching to a fully funded system continues, eight countries in Latin America claim to have either abandoned or are in the process of abandoning their pay-as-you-go systems in favor of fully funded systems.[4] Mexico is one of these eight countries, and it is of particular interest to U.S. analysts because of both its geographical proximity and close relationships with the United States and the similarities of its reform program to what many policymakers and economists advocate for the U.S. system.

The Mexican government claims that it has started a move to a fully funded system. As proof, proponents of the new system point out that since 1997 Mexico has adopted a privately managed defined-contribution system. It is important to emphasize, however, (as is done in Espinosa-Vega and Russell 1999) that a pension system can be privately administered without being fully funded. The new system is seen in some circles as a great accomplishment. Proponents

of the pension reform (for example, Rodriguez 1999 and Sales-Sarrapy, Solís-Soberón, and Villagómez-Amezcua 1998) predict that it will lead to a number of positive future developments: (1) The system will be actuarially balanced. (2) It will increase private (and national) saving. (3) Workers will migrate from the informal to the formal labor market. (4) More workers will be covered by the social security system. (5) The new system will create long-term investment instruments. But before uncorking the bottle of champagne, it is important to ask a few questions. Has Mexico started a migration toward a fully funded system? What are the likely net gains from the Mexican pension reform? Are predictions 1 through 5 likely to materialize?

**Mexico is in dire need of further research to guide it through its decision on whether and how to switch to a fully funded pension system.**

There is voluminous literature on social security systems, both country-specific and general. A survey on this literature is beyond the scope of this article. The objective here instead is to provide a primer on the Mexican pension system and to evaluate it critically. The ultimate intended goal in analyzing the Mexican experience is to illustrate the difficulties in assessing the economic significance of a pension reform. In general the hope is that in the current environment where every other country seeking reform claims to be jumping on the fully funded wagon, this discussion may help to temper expectations.

The article traces some of the official rationales for the reform in Mexico and provides a summary of the new developments leading to it. It reports its operational rules and the critical elements of the new pension system. The article also applies the insight of a companion piece by Espinosa-Vega and Russell (1999) to assess the significance of the changes introduced by the reform. It makes clear that while the reform is likely to bring some benefits, it also has costs (something that has not been emphasized in the existing literature). Finally, it calls for further research to appraise the predictions spelled out above and the net benefits of the reform for the Mexican society. In the end, the Mexican case provides a good case study for those countries that are either considering or have engaged in a pension reform of their own.

## Key Features of Mexico's Old Social Security System

The next sections introduce the most significant features of the old Mexican public pension system as a point of reference for discussing the reform. There have in fact been several pension plans in Mexico. Each of these plans is in turn part of a larger benefits plan. Federal employees' accounts are managed by the Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado (ISSSTE). There is a special fund for the state-owned petroleum-related monopoly, PEMEX. Private-sector workers' accounts have been managed by the government-run Instituto Mexicano del Seguro Social (Mexican Social Security Institute: IMSS). Furthermore, within each of these institutions, health insurance, housing programs, and social security programs are bundled together. Because the first two systems have been left intact, this discussion focuses solely on IMSS and more particularly on the old-age security aspect of IMSS, which is the core of the current Mexican pension reform.

The IMSS started its operation in 1943–44. Its social security chapter was designed to cover four areas: disability, old age, severance, and disability and life insurance (Invalidez, Vejez, Cesantia en Edad Avanzada, y Muerte, or IVCM). As stated in Grandolini and Cerda, "The original IMSS-IVCM can be characterized as a partially funded defined benefit scheme. However, since the very beginning, it operated as a pay-as-you-go scheme as the fund's actuarial reserves were used to finance other social insurance activities, particularly health" (1998, 4).

Restricted to IMSS, the reform affects only the portion of the economically active population working in the formal private sector. The fact that this sector is proportionally smaller than its counterpart in developed economies may be relevant in assessing the macroeconomic impact of the reform. More than 40 percent of Mexico's 33.5 million labor force is outside the formal sector (Judisman 1997), working in what the International Labor Organization (van Ginneken 1998) calls the informal sector: independent workers (excluding professional, administrative, and technical personnel), domestic workers, and workers in small enterprises (with five workers or fewer).[5] Of the total economically active population, social security covers less than one-third. To put it differently, it covers slightly more than half the workers in the formal sector. Therefore, talk about reform is talk about directly affecting only half the formal labor force.

The discussion that follows reviews other relevant features unique to the Mexican pension system. It starts by looking at the eligibility criteria for workers to qualify for benefits (called the admission fee) and the fraction of individuals' working income received upon retirement (the so-called replacement rate).

## Benefit-Eligibility Requirements and the Replacement Rate under the Old System

For retirees the most important aspect of their benefits is the proportion of wages received during their active years that is replaced by their retirement income. This proportion is called the replacement rate. For example, a replacement rate of 100 percent would mean that an individual's annual income would be the same before and after retirement. The concept of replacement rate is significant because eligibility requirements are often quoted in terms of it. It is also important to review the replacement rate and eligibility criteria here to be able to contrast the old system with the new.

For most individuals, wages tend to rise with age. Therefore, it is incorrect to talk about a single replacement rate. Instead, it is customary to discuss the replacement rate with respect to either lifetime-average wage or final salary (and in some cases with respect to an average of a worker's five or ten highest-income years).

In the old regime, the system of old age (and disability) benefits was designed so that a worker became fully eligible to receive benefits after just 500 weeks of contribution. However, benefits did not increase much with additional contribution, as Table 1 illustrates. Moreover, if for some reason a worker stopped contributing for 500 consecutive weeks, he or she would lose all retirement benefits.

**TABLE 1**
**Replacement Rates under Mexico's IMSS System (Percent)**

| Salary at Retirement | Years in the System | |
|---|---|---|
| | 10 | 30 |
| 1 minimum salary | 100 | 100 |
| 6.5 x minimum salary | 23 | 25 |
| 10 x minimum salary | 14 | 16 |

Source: Serrano (1999a)

The numbers presented in Table 1 are instructive. The first row states that for a person earning one minimum salary, the replacement rate would be 100 percent regardless of the number of years the individual contributed to the pension fund. Things are not very different for other participants. A person earning ten times the minimum salary, for example, would get 14 percent of his or her wage replaced after ten years and 16 percent of his annual salary replaced after thirty years. In this case, although the incentive to contribute to the pension fund for more than ten years was not zero, it was minimal. Recent estimates show that 86 percent of current retirees get exactly one minimum salary as the retirement benefit (Sinha 1999a).

Thus there was a fairly low minimum admission fee. The fact that workers qualified for pension after only 500 weeks of work created an incentive to contribute just long enough to become eligible

1. For a detailed description of the key differences between pay-as-you-go and fully funded systems, see a companion article by Espinosa-Vega and Russell (1999).
2. Feldstein's (1974) theoretical analysis suggests that privatization of social security would reduce the distortions that payroll taxes impose on household saving and labor supply decisions. Even in the absence of redistributional considerations or the presence of market imperfections, Feldstein's work, as well as that of his successors, is subject to a qualification shown in Diamond's (1965) theoretical analysis: a mandated pay-as-you-go defined-contribution social security system would improve a country's well-being provided the economy was dynamically inefficient. (Roughly put, a competitive economy is said to be dynamically inefficient if it saves "too much" relative to the social optimum.)

   Imrohoroglu, Imrohoroglu, and Joines (1995) extend Diamond's general equilibrium work by adding potentially more realistic lifetime structure and market imperfections. They are able to show how the replacement rate (the ratio of retirement benefits to preretirement wages) varies according to the market structure and specific parameter assumptions. Because their analysis focuses on economies that are dynamically inefficient, Diamond's result prevails. Abel and others (1989) provide empirical support for the dynamic efficiency of the U.S. economy. Building on their work and expanding Feldstein's analysis to a general equilibrium framework, Kotlikoff (1996) has provided extensive simulation analysis for the U.S. economy that supports Feldstein's conclusion. In a framework that allows for intra- and intergenerational redistribution, these authors show that in a competitive economy privatization of the social security system would—after intragenerational lump-sum transfers if necessary—improve the well-being of the country. A recent example of a serious critique of a fully funded scheme is found in Sinn (2000).
3. See Schwarz and Demirguc-Kunt (1999) for a complete list of countries engaged in pension reform.
4. The countries are Argentina, Bolivia, Chile, Colombia, El Salvador, Mexico, Peru, and Uruguay.
5. In Latin America, more than half of the economically active population work in the informal sector.

and then either drop out to the informal sector or "unregister" with the IMSS and continue working without being officially on the payroll. This awkward eligibility requirement was another factor contributing to the actuarial imbalance of the IMSS-IVCM under the old system, which was funded essentially by a payroll tax. At the same time, because employers were responsible for paying part of this tax, many of them understated the wage rate of workers just to avoid paying the payroll tax.

In addition, the government had relaxed eligibility by, for example, relaxing the age of retirement, by using broader definitions of disability or poor health, and so forth. One manifestation of this problem, which was severe in the Mexican system, is that an increasing number of people were getting a disability pension. Since the middle of the 1980s, the proportion of people drawing a disability pension has stayed at more than 40 percent (see Table 2), a very high figure compared with Organisation for Economic Cooperation and Development (OECD) countries, whose population is generally much older.

As is described below, one can identify two opposing factors affecting the balance of the pension portion of the IMSS fund. Even though the replacement rate appears generous at first glance, it is quoted in terms of the minimum salary. The minimum salary in Mexico at the time of the reform was roughly $24 a day. The World Bank (2000) considers that in developing nations an "adequate" standard of living can be maintained with $40 a day. This low level of disbursement (in combination with the large proportion of young to old people described below) worked to boost the coffers of the IMSS. On the other hand, the low admission fee made it unattractive to stay in the system for more than ten years and thus constituted a strain on the coffers of the system.

At the same time, there were other strains on the retirement account of the IMSS. The Mexican benefit system has historically been tied to the minimum wage (that is, it has always been calculated as a multiple of the minimum wage), which is adjusted only by legislation. Indexing of retirement benefits was first introduced in the Mexican system in 1989, when Congress passed a law stating that for calculation of IMSS benefits the minimum wage would be indexed to the consumer price index. The government thereby increased the benefits of the retired population by indexing benefits to inflation but added to strains on the IMSS because it did not at the same time index revenue to inflation.

In spite of these idiosyncrasies, from its inception the private pension system in Mexico operated with surpluses because of favorable demographic factors. For example, behind these surpluses lay a large base of contributors relative to benefit recipients. However, for most of those years, instead of building reserves these surpluses were used to subsidize IMSS's other programs such as its health insurance component. According to the IMSS, this status quo was sustainable without any changes until the year 2007. However, as the next section illustrates, in recent years Mexico has experienced dramatic changes in mortality rates and demographic trends, changes that would have reduced and even eliminated the surpluses on the IMSS pension accounts.

**TABLE 2**
**IMSS-IVCM Disbursements by Old-Age Retirement and Disability Categories (Percent)**

| Year | Old Age | Disability |
|---|---|---|
| 1981 | 64.95 | 35.05 |
| 1985 | 58.86 | 41.14 |
| 1990 | 56.47 | 43.53 |
| 1994 | 57.01 | 42.99 |

Source: IMSS (1997)

## The Demographic Angle

In recent years Mexico has experienced a significant drop in its fertility and mortality rates, which has led to a relatively rapid aging of its population. For example, the proportion of population above age sixty in France was 5 percent in 1750. Mexico reached the same milestone in 1985. However, by 1985 the proportion of French population older than sixty rose to 15 percent. It took France 235 years to get to that point. Mexico will reach this number by 2025, in only 35 years. France had the opportunity to change its social institutions slowly to cope with the problems associated with population aging. Mexico, on the other hand, has had to expedite its social security reform.

Table 3 presents a clearer picture of how rapidly population changes are occurring in Mexico. The table shows actual population proportions for 1970 and 1990. In addition, it includes projected population proportions in 2010, 2030, and 2050. As the numbers clearly show, over a period of eighty years (between 1970 and 2050), the proportion of population older than sixty rises from 6.13 percent to 24.35 percent.

One reason for such a dramatic change in population structure is a rapid decline in fertility rates. In 1970 the mean fertility rate of women was 6.5 children per lifetime. This figure is projected to fall

**TABLE 3 Actual and Projected Changes in Age Distribution, 1970–2050**

| Age | 1970 | 1990 | 2010 | 2030 | 2050 |
|---|---|---|---|---|---|
| 0–4 | 18.59 | 13.20 | 9.38 | 7.38 | 6.43 |
| 5–9 | 15.14 | 12.65 | 9.47 | 7.31 | 6.42 |
| 10–14 | 12.77 | 12.70 | 9.59 | 7.37 | 6.45 |
| 15–19 | 10.38 | 12.18 | 9.39 | 7.33 | 6.43 |
| 20–24 | 8.22 | 9.86 | 8.80 | 7.19 | 6.30 |
| 25–29 | 6.78 | 7.97 | 8.32 | 7.21 | 6.18 |
| 30–34 | 5.53 | 6.72 | 8.28 | 7.26 | 6.20 |
| 35–39 | 4.69 | 5.49 | 7.99 | 7.16 | 6.22 |
| 40–44 | 3.90 | 4.40 | 6.60 | 6.86 | 6.24 |
| 45–49 | 3.33 | 3.62 | 5.38 | 6.54 | 6.31 |
| 50–54 | 2.44 | 2.92 | 4.51 | 6.48 | 6.33 |
| 55–59 | 2.13 | 2.41 | 3.60 | 6.14 | 6.14 |
| 60–64 | 1.87 | 1.91 | 2.76 | 4.89 | 5.71 |
| 65–69 | 1.53 | 1.50 | 2.14 | 3.79 | 5.21 |
| 70–74 | 1.20 | 0.97 | 1.56 | 2.92 | 4.80 |
| 75–79 | 0.85 | 0.71 | 1.11 | 2.06 | 4.08 |
| 80+ | 0.68 | 0.77 | 1.11 | 2.11 | 4.55 |
| Total | 100 | 100 | 100 | 100 | 100 |

Source: Data from United Nations (1998, table 3)

to 2.1 by 2050. At the same time, the infant mortality rate fell from sixty-nine per thousand live births to eleven per thousand live births. These two trends have opposite effects with the decline in fertility leading to fewer people entering the workforce and the improved infant mortality rate somewhat alleviating this problem. At the same time, the mortality rate of people in higher age groups has also fallen, contributing further to the aging of the population structure.

All these changes can be summarized in what is called the dependency ratio of the population. The dependency ratio is usually defined as the number of people in 0–14 and 65+ age groups (the dependent group) divided by the number of people in the 15–64 age group (because the labor force usually consists of the latter age group).

Over the eighty-year period from 1970 to 2050, the dependency ratio changes dramatically (from 1.03 to 0.52) in the first forty years. As Table 4 shows, it drops from 1.03 to 0.52. Thus, the number of people dependent on the working-age population by 2010 will have fallen by 50 percent. Then it is projected to rise somewhat. This rise is somewhat deceptive, however, hiding the composition of the dependent population. The change in composition of the dependent population is evident in the ratio of old to young in the population, which moves from 9 percent to 97 percent over the total period. This scale of change in age composition has been witnessed by very few countries over such a short time.

**TABLE 4**
**Actual and Projected Dependency Ratios, 1970–2050**

| Indicator | 1970 | 2010 | 2050 |
|---|---|---|---|
| Dependency Ratio | 1.03 | 0.52 | 0.61 |
| Old-Young Ratio | 0.09 | 0.21 | 0.97 |

Note: The dependency ratio is the number of people in age groups 0–14 and 65+ divided by the number of people in the 15–64 age group. The old-young ratio is the number of people in the 65+ age group divided by the number of people aged 0–14.

Source: Atlanta Fed calculation using data from United Nations (1998)

In view of these demographic changes, policymakers have had to face a pressing question: How

onerous would maintaining the status quo be? The discussion now turns to this question.

## Estimating the Actuarial Imbalance

So far, the discussion has identified and described the strains to the Mexican private pension system without actually reporting what it would have cost the government to maintain the status quo under IMSS-IVCM. Without trying to evaluate their accuracy, this section reports three such estimates.

Table 5 contains a projection attributed to IMSS by Grandolini and Cerda (1998). The table reports that the present value of IMSS commitments through 2058 as of December 31, 1994, (the year Congress started to consider a second reform) was 142 percent of the 1994 gross domestic product (GDP) present-value deficit.

To get a sense of how the time path of actuarial deficit would play out had there been no changes in the system, IMSS itself calculated the projected deficit. The IMSS figures are reproduced here as Table 6. The table shows that the IMSS would have run a surplus until 2005 (a positive number in the table indicates a surplus) had the old system not been changed. Therefore, the situation in Mexico was not like that of Argentina or Uruguay (where the governments were already filling up the deficits of their pay-as-you-go pension systems with current government budgetary resources). On the other hand, after 2020 the deficit would have mounted rapidly.

Sales-Sarrapy, Solís-Soberón, and Villagómez-Amezcua (1998) present an alternative estimate of the cost of maintaining the IMSS-IVCM status quo. The least costly of their scenarios has the cost going from 1.55 percent of GDP in 1997 to 3.59 percent in 2022 and 6.69 percent in 2047.

Why do these estimates of the deficits differ? For example, according to the IMSS figures, for 1997 there was a surplus in the pension fund. On the other hand, Sales-Sarrapy, Solís-Soberón, and Villagómez-Amezcua (1998) report a deficit for 1997. Given the information provided by the different authors, it is impossible to identify explicitly the reason for most of these differences, and it is therefore impossible to adequately compare the different estimates.

An additional challenge is that not all estimates consider the same concepts. The concept of implicit pension debt measures the stock of debt today. If all the taxes to finance the pay-as-you-go system are set to zero, the implicit pension debt shows how much the government owes (implicitly) to the current generation as of today. There is no liability for future generations in this calculation. The calculation is the exact analog of government debt with one difference: explicit government debt does not depend on the mortality experience of the current generation. On the other hand, implicit pension debt does because most governments promise pensions for widows (and sometimes to other dependents).

This concept should be contrasted with that of the present value of cash flow deficits. As the name suggests, cash flow deficits are calculated as the difference between expected contributions at every future date, which in most cases represents a deficit. Then, the present value of the stream of numbers is calculated. If contribution rates and benefits rates do not change but the underlying demographics do, the deficit will be altered. Specifically, aging of the population will make deficits worse. The period over which the deficit is calculated also matters. The larger the period, of course, the bigger the deficit.

The issue is further complicated because authors may not explicitly identify the concepts with which they are working. For example, the Grandolini and Cerda (1998) and Sales-Sarrapy, Solís-Soberón, and Villagómez-Amezcua (1998) studies do not always clarify whether they are talking about implicit pension debts or cash flow deficits.

Additional problems arise from the fact that there is no universal standard for the discount rate chosen to calculate the present value. For example, Grandolini and Cerda (1998) chose to use a 3 percent discount rate. The advantage of Table 6 is that it allows avoiding taking an arbitrary discount rate. Instead of all the numbers being lumped by being added up, they remain a vector of values. The significance of such confusions is that they can lead to vastly different conclusions (see Sinha forthcoming, chap. 3, especially table 3.33.) Nonetheless, without attempting to homogenize the different estimates of the cost of maintaining the status quo under IMSS-IVCM, it is clear that, according to these studies, maintaining the status quo would have been very costly for the country.

## The New Mexican Social Security System

In December 1995, the Mexican Congress passed the new Social Security Law (Ley de Seguro Social), paving the way for the current system. A second set of laws (Ley de los Sistemas de Ahorro para el Retiro) was passed in April 1996. These laws allowed privatized management of the country's pension system. They approved operation of investment management companies (Administradores de Fondos de Ahorro, or AFOREs) to manage individual retirement funds (Sociedades de Inversion

**TABLE 5**
**Present Value of Future Pension Deficits (in Billions of Pesos) as of December 31, 1994**

| Assets | | Liabilities | |
|---|---|---|---|
| Reserves | 3.25 | PV of old pension | 96.93 |
| PV of future contributions | 683.67 | PV of future liability | 2,390.61 |
| (Affiliates now) | 179.74 | (This generation) | 1,017.40 |
| (Future generations) | 503.93 | (Future generations) | 1,373.21 |
| Total | 683.92 | Total | 2,487.54 |

Source: Grandolini and Cerda (1998)

**TABLE 6 Actuarial Deficit Projection of IMSS If the Old System Had Continued**

| Year | Millions of 1994 Pesos | Percent of 1996 GDP |
|---|---|---|
| 2000 | 9,916 | 0.39 |
| 2005 | 667 | 0.03 |
| 2010 | –23,407 | –0.93 |
| 2015 | –63,950 | –2.55 |
| 2020 | –122,827 | –4.89 |
| 2025 | –200,741 | –8.00 |
| 2030 | –264,501 | –10.54 |

Note: Some of the estimates that went into computing Tables 5 and 6 include (1) Demographics: sizes of workers and retirees of every generation in the future. These numbers will in turn depend on fertility and mortality projections (ignoring migration). (2) Estimates of growth rates of real wages in the future. (3) Retirement pattern of the elderly in the future. (4) Participation rate of women and other part-time workers in the labor force. (5) Proportion of economically active population participating in the formal sector. (6) Inflation rate projection. Under the old regime, the benefits are calculated on the basis of the average nominal salary of the last five working years. It also required a choice of a discount rate to convert these figures to a single number. Although the authors reveal that these are partial equilibrium computations the exact methodology is not spelled out in the document.

Source: IMSS (1997, table 18)

Especializadas en Fondos para el Retiro, or SIEFOREs). In addition, the Mexican government set up a separate division to oversee all activities of the AFOREs: Comisión Nacional del Sistema de Ahorro para el Retiro (CONSAR). To clarify the roles of the AFOREs, CONSAR has set out general rules of operation for the companies (see Banco de Mexico 1996).

The stated objectives of AFOREs include the following: (1) To open, administer, and manage the individual retirement accounts in agreement with provisions in social security laws. Regarding housing-promotion subaccounts, the AFOREs will register each worker's contributions and the interest paid thereon, using information provided by social security institutions.[6] (2) To receive from social security institutions the contributions made, in accordance with the law, by the government, employers, and workers, as well as voluntary contributions by workers and employers. (3) To itemize the amounts received periodically from social security institutions and deposit them into each worker's individual

6. The housing subaccount requires a contribution of 5 percent of wages. This amount is substantial (the retirement contribution is 6.5 percent of wages). In the past, this housing subaccount has earned a negative real rate of return. All future estimates assume that it will earn a zero real rate of return. One interesting question is, Why is the government so keen on getting the house in order for the retirement account but not touch the housing subaccount?

retirement account as with the returns obtained on the investment of these funds. (4) To provide administrative services to mutual investment funds (the SIEFOREs). These are direct subsidiaries of the AFORES. In fact, at present, each AFORE is allowed to have one SIEFORE.

**Contribution Structure.** The contribution structure of the new system is as follows: Each individual pays a compulsory 6.5 percent of wages into an individual retirement account. The government contributes a "social quota" (called cuota social) of 5.5 percent of minimum wage (regardless of the wage rate of the worker). This social quota is funded from the government's general revenue every year; thus the funding mechanism is taxes on the current generation of workers. In addition, workers must contribute 5 percent to a housing subaccount (INFONAVIT) that will be consolidated with the AFORE account upon retirement. Also, 4 percent of wages go to IMSS for disability and survivors insurance. Workers can also make additional voluntary contributions. The AFOREs started to collect compulsory and voluntary contributions in February 1997. Contribution to the new system became compulsory for all private-sector workers in September 1997.

AFOREs are allowed to charge management fees either as a percentage of contribution, a percentage of value accumulated, or any combination thereof. Most AFOREs charge fees as a percentage of contribution. All are required to inform affiliates about their accounts at least once a year with statements that include information about accumulated value, contributions during the year, and any charges the account has incurred.

**Contribution Requirements: A Comparison.** In order to gain some perspective on the differences between the required contributions and on eligibility requirements under the old and the new social security systems, the following information is provided. The box on page 19 is a compilation of information provided by CONSAR, IMSS, and SHCP (Secretaría de Hacienda y Crédito Público) and reported in Sales-Sarrapy, Solís-Soberón, and Villagómez-Amezcua (1998, 146) and Grandolini and Cerda (1998, 13).

The box allows identifying at a glance some of the idiosyncratic features of the old system mentioned above that have been eliminated. For one thing, the minimum ten-year contribution necessary to qualify for retirement benefits has been replaced by a minimum twenty-five-year contribution. Also, because there is only a minimum defined benefit under the new regime, the asymmetric inflation-indexing problem described above should be eliminated. And because the notion of a social security surplus has been eliminated, the funds can no longer be a source of subsidy for other IMSS activities. At the same time, because the IVCM has been separated from the health care and maternity benefits provided by IMSS, deficits in these areas will be directly reflected in government deficits.

**Issues Involving the Fund Managers.** Workers can choose any AFORE for contribution. Once an AFORE is chosen, no change can be made for one year, though it is possible to choose a different AFORE every year without any financial penalty. In Mexico, fund-hopping has been very low. In 1999, less than 0.01 percent of workers changed funds. This stability stands in sharp contrast with Chile, where fund-hopping has exceeded 25 percent per year.

By the end of 1997 CONSAR had licensed seventeen AFOREs (listed in Table 7). Some of the AFOREs are fully owned by Mexican companies, and others are partly owned by foreign companies. For example, AFORE Bancomer is 51 percent owned by the second-largest banking group in Mexico, and the remaining 49 percent is owned by Aetna, one of the largest insurance companies in the United States. Garante has a particularly interesting ownership structure with majority shareholding by a Mexican group, part ownership by Citibank, and part by a pension fund from Chile, AFP Habitat. Ownership structure of Siglo XXI is also notable: half of it is owned by the IMSS, the government organization that continues to run health care and disability and death insurance for the entire system.

Three of the AFOREs established in 1997 have merged with others. Confia bought Atlantico, Santander bought Genesis, and Profuturo bought Previnter. Consequently, as of August 1999, fourteen AFOREs are left in the market. All of these mergers had to be approved by CONSAR.

**Market Share.** There were two very distinct waves of membership in the new social security scheme. The first was the initial rapid expansion until the number of affiliates hit around 10,000,000 within a span of ten months (see Chart 1). Then came a second, slower stage of expansion over the next fourteen months. At the end of August 1999, about 14,900,000 workers had signed up for one AFORE or another.

It should be noted that of the approximately fifteen million workers who belonged to some AFORE in August 1999, about 87 percent are active contributors. The fact that an individual signs up and becomes an affiliate does not necessarily mean that

**TABLE 7 AFOREs Authorized by CONSAR, 1997**

| AFORE | Main Shareholders with Percentage Holding |
|---|---|
| Atlántico Promex | Banca Promex, 50; Banco del Atlántico, 50 |
| Banamex | Grupo Financiero Banamex-Accival, 100 |
| Bancomer | Grupo Financiero Bancomer, 51; Aetna Internacional, Inc., 49 |
| Bancrecer-Dresdner | Grupo Financiero Bancrecer, 51; Dresdner Pension Fund Holdings, 44; Allianz México, S.A., 5 |
| Bital | Grupo Financiero Bital, 51; ING America Insurance Holding Inc., 49 |
| Capitaliza | General Electric Capital Assurance Co., 100 |
| Confia-Principal | Abaco Grupo Financiero, 51; Principal International, 49 |
| Garante | Grupo Financiero Serfín, 51; Grupo Financiero Citibank, 40; Hábitat Desarrollo Internacional, 9 |
| Génesis | Seguros Génesis, S.A., 100 |
| Inbursa | Grupo Financiero Inbursa, 100 |
| Previnter | Boston AIG Company, 90; Bank of Nova Scotia, 10 |
| Profuturo GNP | Grupo Nacional Provincial, 51; Banco Bilbao Vizcaya-México, S.A., 25; Provida Internacional, S.A., 24 |
| Santander Mexicano | Grupo Financiero Inverméxico, 75; Santander Investment, S.A., 25 |
| Siglo XXI | Instituto Mexicano del Seguro Social, 50; IXE Grupo Financiero, 50 |
| Sólida Banorte | Grupo Financiero Banorte, 99 |
| Tepeyac | Seguros Tepeyac, 99 |
| Zurich | Zurich Vida, Compañía de Seguros, 77; Gabriel Monterrubio Guasque, 10 |

Note: No mention is made of shareholders with equity participation under 5 percent of the total capital of the respective AFORE.

Source: Banco de Mexico (1997)

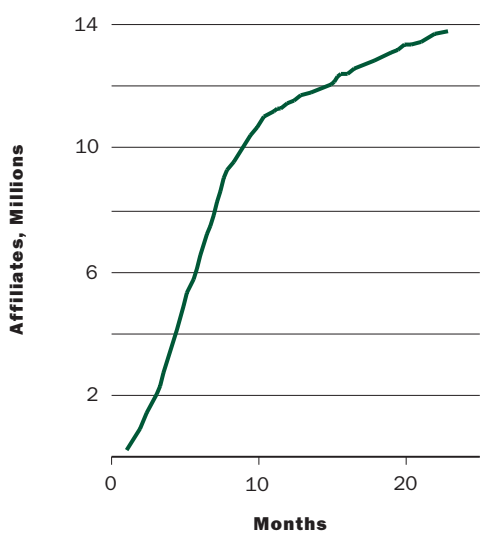
he or she will contribute to the system regularly. In addition, each may have more than one account, inflating the number of affiliates. SAR (Sistema de Ahorro para el Retiro) accounts provide one classic example: by the end of 1995, there were 65 million accounts in SAR but less than twelve million workers in the formal sector.

The amount of contributions in the system has also increased steadily. Between July 1997 and July 1998, investment in the system equaled US$3 billion (at an exchange rate of 10 pesos per U.S. dollar as of January 1999). Over the next seven months (July 1998 to January 1999), investment grew another US$3 billion. If this trend continues, in twenty-five years AFOREs will hold an amount equal to 40 percent of GDP (assuming a real GDP growth rate of 2 percent a year and real rate of return of funds at 6 percent a year).

Table 8 presents a summary of compulsory and voluntary contributions to the existing AFOREs as of the end of 1998. As is evident, the market is highly concentrated, a feature common to other Latin American countries such as Chile and Argentina (see Queisser 1998).

CONSAR has explicitly prohibited any AFORE from holding more than 17 percent of market share

**CHART 1**
**Systemwide Take-Up Rates of AFOREs through January 1999**




Source: CONSAR (1999)

in terms of the number of affiliates. However, it does not restrict market share in terms of total value of assets in portfolios. For example, an AFORE may have 20 percent market share in terms of its portfolio's value but less than 17 percent in terms of the number of affiliates. The value of investments in both Banamex and Bancomer has exceeded 17 percent during most of the past two years, for instance. As Table 8 shows, the level of concentration in terms of total investment portfolio is far higher than in terms of the number of affiliates: six companies have around 77 percent of the market share in terms of investments. Contrary to what the regulation intended, with more market consolidation, this proportion is likely to rise in the future.

**The Portfolios of the Fund Managers.** An integral component of any pension system is the composition of the portfolio held on behalf of its contributors. As Table 9 shows, AFOREs' portfolios are heavily concentrated in government bonds.

This portfolio composition results from CONSAR's stipulation that a minimum of 51 percent of an AFORE portfolio be held in the form of inflation-indexed bonds and at least 65 percent in assets with a maturity of no more than 183 days. CONSAR's reasons for this portfolio requirement are to build trust in the system and avoid volatility in the portfolio (see CONSAR 1999). On January 31, 1999, more than 66 percent of AFOREs' portfolios were in inflation-indexed bonds (called BONDE91 and UDIBONOS). Another 22 percent were in CETES (Mexican Treasury bills). The average maturity of investment portfolios is 111 days, well below CONSAR's cap.

CONSAR specifies that an AFORE can hold up to 35 percent of its portfolio (disposicion quinta, 5) as private debts (CONSAR 1997). Given this range, why do private debt holdings amount to only 2.83 percent of all portfolio assets? CONSAR's qualifications on the type of debt that can be included probably account for the low figure. Regulations specify that only private short-term debt meeting Standard and Poor's mxA-3 grade or equivalent and long-term private debt meeting Standard and Poor's mxAA+/mxAA grade would make the cut (CONSAR 1997, chap. 3 and app. A), and only a very small fraction of Mexican private debt meets these eligibility requirements.

**Recent Proposals for Portfolio Changes.** Recently there has been criticism about the need for Mexico to move forward with privatizing its pension system (for example, Rodriguez 1999). Specifically, criticism has focused on the fact that the AFOREs' portfolios consist mostly of government debt. Even though, as explained in Espinosa-Vega and Russell (1999), there would be no guarantee that the new system would be fully funded if the government relaxed its high government-debt requirement for AFOREs' portfolios, questions remain about the economic impact of allowing AFOREs more flexibility in the composition of their portfolios.[7] Recently, the Mexican federal government has been considering a proposal to allow AFOREs greater flexibility on how retirement savings are invested. CONSAR has proposed to increase investment options for the fund administrators, including the right to buy debt sold by Mexican corporations in overseas markets. It is not hard to foresee a trend in the direction of allowing AFOREs to have larger holdings of stocks. As explained in the next paragraph, this move, together with the issuance of inflation-indexed long-term government bonds, may represent net gains for the generation of active workers when they retire.

In the shift from defined benefits under the old system to defined contributions under the new system, risk is shifted from workers' active years to their years of retirement. The reason for this shift involves risk created by the possibility of increases in the inflation rate. Under Mexico's old defined-benefits social security system, the nominal (peso) value of the benefits was indexed to the inflation rate. Thus, workers did not have to fear that the purchasing power of their social security benefits would be reduced by higher inflation. However, during periods when the government had budget prob-

**TABLE 8 Money Invested and Number of Affiliates in AFOREs**

| Fund | Amount in Pesos[a] | Percent | Affiliates[b] |
|---|---|---|---|
| Banamex | 10,209,603,059 | 18.16 | 1,742,930 |
| Bancomer | 13,183,450,811 | 23.45 | 2,364,074 |
| Bancrecer | 1,981,244,448 | 3.52 | 619,789 |
| Banorte | 2,628,350,064 | 4.68 | 1,260,762 |
| Bital | 4,772,054,894 | 8.49 | 1,499,758 |
| Confía | 1,027,806,121 | 1.83 | 332,999 |
| Garante | 4,824,234,812 | 8.58 | 1,633,528 |
| Génesis | 302,320,757 | 0.54 | 140,957 |
| Inbursa | 5,072,806,294 | 9.02 | 378,376 |
| Profuturo | 5,155,845,595 | 9.17 | 1,998,211 |
| Santander | 3,534,727,403 | 6.29 | 2,026,656 |
| Siglo XXI | 3,041,053,960 | 5.41 | 462,473 |
| Tepeyac | 285,687,001 | 0.51 | 228,621 |
| Zurich | 199,941,483 | 0.36 | 185,576 |
| Total | 56,219,126,705 | 100 | 14,874,710 |

[a]As of the end of 1998
[b]As of the end of August 1999
Source: CONSAR (1999)

**TABLE 9 Asset Allocation of AFOREs at the End of 1998**

| Type | Amount in Pesos | Percent |
|---|---|---|
| Nominal government | 13,699,239,431 | 22.54 |
| Real government | 40,139,931,258 | 66.04 |
| Repurchase agreements | 1,810,442,216 | 2.98 |
| Private papers | 1,722,866,789 | 2.83 |
| Deposit in Banco de Mexico | 3,404,749,272 | 5.60 |
| Total | 60,777,228,966 | 100 |

Note: Nominal government means government bonds denominated in nominal terms. Real government means government bonds denominated in real (inflation-adjusted) terms.

Source: CONSAR (1999)

lems—for example, if weak performance of the economy reduced tax revenues or high interest rates increased the burden of debt service—the government had to increase taxes or borrow to maintain the level of benefits.

Under the new system, an increase in the inflation rate will reduce the purchasing power of the government or private bonds with fixed nominal values held by the social security system. This inflation will also reduce the purchasing power of benefits paid to retirees. The government may be tempted to take advantage of this fact and increase the inflation rate during periods when it has budget problems, rather than increasing taxes or borrowing. Thus, the new

7. For a fuller explanation of the related issues, see Espinosa-Vega and Russell (1999).

system may increase the risk facing retirees, but it will reduce the risk facing active workers.

On the other hand, there are reasons to expect that the amount of inflation risk facing retirees under the new system may not be very large. First, some of the assets held by the Mexican social security system will consist of stock, and the rate of return on stock tends to increase when the inflation rate increases so that some of the reduction in purchasing power from retiree benefits from government or private bonds will be offset. Second, most of the bonds held by the system are likely to be short-term bonds. The long-term bond market in Mexico currently has a low volume of issues and little trading, and it will probably stay that way unless or until the government convinces the public that it is unlikely to indulge in high inflation. As the inflation rate increases, maturing short-term bonds can be replaced with new short-term bonds yielding higher interest rates, so the losses from inflation will be limited. Finally, the interest rates on many of the longer-term bonds purchased by the system may be indexed to the inflation rate. The Mexican government has recently begun to issue indexed bonds in substantial quantities.

**Transactions Costs and Commissions.** A key source of dissatisfaction and confusion with most newly privatized pension systems is the fees charged by the fund managers (also called commissions or costs of transactions). The main concerns are that the management fees are high, that the fees are lumped with insurance premiums for the life insurance component of the pension plans (see Sinha forthcoming, chap. 3), that the management fees are obfuscated because in most cases they are presented as a fraction of a worker's salary, and that sometimes it is not clear whether the commissions are expressed as a proportion of flow into the fund on a yearly basis or as a proportion of balance in the fund at a given point in time. As pointed out by Diamond (1998), some of these concerns are widespread. For that reason, this section takes a close look at the commission structure for the Mexican pension system.

Table 10 gives the details of the commissions charged by the 17 AFOREs that started out in 1997. The three of them that have since merged with others have assumed the names of the companies under which they are operating.

Most of the commissions charged apply to the flow of contributions. However, some companies charge on the balance in the fund as well as on flows. One company (Inbursa) charges commissions exclusively on the real (inflation-adjusted) rate of return of the fund. (Inbursa charges no fee if the real rate of return is not positive.) In addition to the different fee structures, the way charges are expressed is somewhat misleading because they are expressed as a percentage of wages and not as a percentage of contribution every year.[8] For example, 1.7 percent charges on a person contributing to the system a mandatory 6.5 percent of her $100.00 wage will be effectively paying a $26.15 (26.15 percent) commission charge.

Given that it is somewhat difficult to compare charges across different AFOREs using the figures in Table 10, CONSAR has worked out commission "equivalents." The idea is that all charges are converted to a charge only on flow of funds. CONSAR publishes these estimates, shown here in Table 11. The results assume that the real rate of return is 5 percent (with no inflation), the charges are applied to a person with three times the minimum wage, and that person has the same income throughout life.

Even after the conversion into equivalent charges on flow, comparisons of different AFOREs are difficult. Among other things, the commission charged is effectively a function of years in the system. It is also a function of factors such as the expected rate of return and the level of wages (some of these results can be seen in Sinha, Martinez, and Barrios-Muñoz 1999). In addition, as Table 11 shows, in some cases (Siglo XXI, Tepeyac, Profuturo, Santander) charges actually go up monotonically without falling over time for the same AFORE. The reason is simple. These companies charge fees on contributions as well as on the account balance. As time passes, more money is accumulated in the accounts, and the bite of charges on the account balance gets bigger.

## Assessing the Reform

The focus of the analysis thus far has been the pension system of the formal private sector in Mexico. The discussion has examined the factors that have accounted for surpluses on this pension system and the strains that, if unchanged, would transform these surpluses into deficits in the near term. It also describes key aspects of the new pension system such as the new managers of the system (AFOREs), their portfolios, and recent proposals to modify their portfolios and looks at the new eligibility criteria, the "admission fee," and the replacement rate under the new system. The next sections establish what it would take to assess the significance of the reform and identify a research agenda.

Most economists agree that the old Mexican pension system was a pay-as-you-go system. Many of the predictions about the state of the economy under the new pension system (for example, a sharp increase in national saving) have the economy resembling one

**TABLE 10 Fee Structure of AFOREs**

| AFOREs | Charges on Flow Each Year (Percent of Wages) | Charge on Account Balance (Percent) | Charge on Real Rate of Return (Percent) |
|---|---|---|---|
| Atlántico Promex | 1.40 | | 20.00 |
| Banamex | | | |
| 1997 | 0.20 | | |
| January 1998 | 0.85 | | |
| March 1998 onward | 1.70 | | |
| Bancomer | 1.70 | | |
| Bancrecer-Dresdner | 1.60 | | |
| Banorte | 1.00 | 1.50 | |
| Bital | 1.68 | | |
| Capitaliza | 1.60 | | |
| Confia Principal | 0.90 | 1.00 | |
| Garante | 1.68 | | |
| Génesis | 1.65 | | |
| Inbursa | | | 33.00 |
| Previnter | 1.55 | | |
| Profuturo GNP | 1.70 | 0.50 | |
| Santander | 1.70 | 1.00 | |
| Siglo XXI | 1.50 | 0.20 | |
| Tepeyac | 1.17 | 1.00 | |
| Zurich | 0.95 | Variable | |

Note: In addition, Bancomer, Banamex, Bital, Garante, and Génesis have discounts for people who stay with their funds for long periods of time. These are not shown in the table above.

Source: CONSAR (1999)

under a fully funded system. Thus, the old and new social security systems seem to be two radically different systems. An important question to ask, however, is whether the new system is truly a fully funded system or simply represents a change in the form of the pay-as-you-go system.

**Fully Funded: Is It There Yet?**[9] There are a number of ways to engage in genuine reform, that is, to go from a given modality of a pay-as-you-go system to a fully funded system. However, choosing the way to reform is not trivial. Each alternative social security scheme implies different costs for different generations of workers, and their implementation thus can be subject to the forces of political discourse. The simplest way to carry out a genuine reform would be to have the current workers pay (on top of their regular pension contributions) for the benefits of the current generation of retirees.

8. The following illustrates the difference: Expressing the commission in terms of a fraction ($z$) of a person's wage ($w$) would mean that her total commission payments $t = z \times w$. The total commission payments by an individual who is required to contribute a fraction ($ss$) of her wage, subject to a commission ($cc$) on her contribution would be $t = ss \times w \times cc$. This means that one can extract the effective commission fee ($cc$) by noticing that $cc = z \times w/(ss \times w)$.
9. This section relies heavily on Espinosa-Vega and Russell (1999). Readers should consult that companion article for a fuller understanding of the issues.

**TABLE 11 Equivalence of Commissions (Percentage of Wages)**

| Fund | One Year | Two Years | Five Years | Ten Years | Twenty Years |
|---|---|---|---|---|---|
| Banamex | 1.70 | 1.70 | 1.69 | 1.65 | 1.58 |
| Bancomer | 1.68 | 1.68 | 1.66 | 1.65 | 1.64 |
| Bancrecer | 1.60 | 1.60 | 1.60 | 1.57 | 1.51 |
| Banorte | 1.16 | 1.18 | 1.23 | 1.29 | 1.44 |
| Bital | 1.68 | 1.68 | 1.68 | 1.64 | 1.61 |
| Garante | 1.68 | 1.68 | 1.68 | 1.68 | 1.68 |
| Génesis | 1.65 | 1.65 | 1.65 | 1.65 | 1.65 |
| Inbursa | 0.36 | 0.43 | 0.64 | 1.00 | 1.73 |
| Principal | 1.43 | 1.43 | 1.44 | 1.44 | 1.41 |
| Profuturo | 1.75 | 1.76 | 1.79 | 1.84 | 1.95 |
| Santander | 1.81 | 1.82 | 1.86 | 1.98 | 2.19 |
| Siglo XXI | 1.52 | 1.53 | 1.54 | 1.56 | 1.60 |
| Tepeyac | 1.28 | 1.30 | 1.34 | 1.47 | 1.69 |
| Zurich | 1.09 | 1.09 | 1.14 | 1.19 | 1.14 |

Source: CONSAR (1999)

Clearly, though, this approach would place an unbearable burden on current workers. An alternative would be to issue debt to pay off the current retirees at the time of the reform and then retire the debt, through time, by taxing the current and future workers for a number of years. Under either scenario, the government's actions at the beginning of the transition process would be the same. Bonds must be issued to obtain funds needed for social security payments to current and near-future retirees.

The actions that will distinguish a transition to a fully funded system from a transition to a pay-as-you-go system of the bond/tax-or-transfer type will occur in the future. If the government switches to a fully funded system, then over the next few generations it must collect enough revenue, via new taxes, to retire the aforementioned bonds. If it is switching to a pay-as-you-go, bond/tax-or-transfer system, however, then it may not have to change its total social security tax collections because it will roll the bonds over indefinitely without retiring any of them.[10]

How can it be known today whether the Mexican government will retire the bonds in the future? That is to say, how can the government's intentions to switch to a fully funded social security system be known at this point? Although the government has announced that it does plan to switch to a fully funded system, it has not announced any plans to increase future taxes and it has not announced any schedule for retiring the bonds. Even if the government did make such announcements, how credible would they be? Future Mexican governments might feel free to ignore them, either by explicitly reversing the decision to retire the debt or by postponing the beginning of the debt-retirement process. Future governments would have plenty of incentive to not follow through. Beginning the bond-retirement process would require increasing taxes, a move likely to be opposed by the voting public.

Viewed in this light, there are good reasons to question the likelihood of Mexico's ultimate success in switching to a fully funded social security system as well as the motivations for the reform. On the one hand, the government may wish to get the credit for initiating a switch to a fully funded system—a system, which, in the view of most economists, would be better for Mexico in the long run. On the other hand, the government may be slow to take any concrete steps to begin the transition to such a system because, as mentioned above, they would be politically costly in the short run.[11] It seems more likely, then, that the switch will turn out to be one from a tax-transfer pay-as-you-go system to a bond-based pay-as-you-go system.

**Why Change Systems?** A switch of this sort may have some significant economic effects, but perhaps,

more importantly, it creates the appearance of reform. It does so in several different ways. First, since switching to a bond-based system could (but does not necessarily) represent the very first step in a transition to a fully funded system, this switch allows the government to claim that it has begun the transition process. Second, the switch to a bond-based system allows the government to privatize a number of aspects of the administration of the social security system—a step that might have some benefits in its own right and that many people are likely to misinterpret as representing more effectual reform. Third, all the idiosyncratic features of the old system discussed above represent economic distortions for the decisions of firms and workers in the system. And, as mentioned earlier, other than demographics, it is these idiosyncratic features that would have contributed to the actuarial imbalance of the Mexican pay-as-you-go system. The adoption of a pay-as-you-go system does not require indexing benefits against inflation while leaving contributions unchanged, awkward qualification criteria (lax eligibility requirements), or allowing the government to use the system as an outright source of government revenue. Elimination of these features could go a long way in improving the operation of the old pay-as-you-go system and solving its actuarial imbalance problems.

**Asymmetric Indexation and the Admission Fee.** Indexing benefits against inflation while leaving contributions unchanged insulates social security recipients, but it represents higher taxes in other sectors or a higher government deficit. This asymmetric indexation became a serious problem when inflation rose to triple digits in 1994–95. This problem will be avoided by the move to defined contributions.

As discussed above, the low admission fee under the old system created the incentive to stop pension contributions altogether after the initial 500 weeks required for becoming eligible for retirement benefits. There is also clear evidence that many employers understated the wage rate of workers just to avoid paying the payroll tax. Under this admission scheme, contributing workers in the formal sector subsidized those in the informal sector that had stopped contributing once they had paid their low admission fee. Under the new system, a minimum twenty-five year contribution is required to qualify for any benefits, and at retirement a worker gets the market return to his contributions. This tightening of the eligibility requirements will eliminate the strain that the old requirements had on the IMSS pension accounts.

**Management Fees.** The new system introduces a feature only implicit in the old system: transaction costs (or commissions) imposed by AFOREs on their contributors. It is not straightforward to compare the transaction fees charged by the different AFOREs. However, careful review (see Sinha 1999a) reveals that most charges (by private pension companies in other Latin American countries) are in the order of 20–25 percent of contributions—ten times higher than charges in defined-contribution plans of Singapore or Malaysia. The usual defense offered by proponents of Mexico's system has been that (1) Chile has a similar cost structure and (2) mutual funds in the United States have similar cost structures. In fact, CONSAR has recently used the data in Table 12 to argue that the commission in Mexico is on the average lower than in other Latin American countries, a claim originally made by Solís-Soberón (1997).

There are potential problems with reaching such a conclusion, however, as comparing Chile and Mexico's commission structures illustrates. Table 12 states commission as a percentage of covered pay, not as a percentage of contribution. In Chile each worker contributes 10 percent of the salary into the system whereas in Mexico the contribution is only 6.5 percent of salary. Thus, as a percentage of contribution, average charges in Mexico would be more than 29 percent (1.919/6.5) while in Chile it is under 23 percent (2.291/10). Another factor is that the government contribution to the system has been ignored. It is difficult to evaluate the government contribution because it varies with the wage rate of the worker. For an average worker earning three times the minimum salary, it amounts to 1.83 percent of the salary. If that factor is added to 6.5 percent, the total is 8.33 percent. Computed this way, the commission charges amount to 23 percent (1.919/8.33), exactly the same as charges imposed in Chile. This complexity illustrates the need for a careful review of the claim that Mexico has managed to reduce the charges that Chile could not.

In contrasting transaction costs of pensions in Mexico with those charged by mutual funds in the United States, the following has to be considered. First, charges for the majority of mutual funds in the United States are on the order of 10 percent (of contribution) and not 25 percent (see Mitchell

10. The government will have to pay the interest on the bonds, but it can do so without increasing its social security tax collections.
11. Cooley and Soares (1999) discuss how a pay-as-you-go system may in fact represent the rational politico-economic outcome in a democratic regime.

**TABLE 12**
**Commission Structure Comparison**

| Country | Percent of Covered Pay |
|---|---|
| Argentina | 2.410 |
| Chile | 2.291 |
| Mexico | 1.919 |
| Peru | 2.294 |
| Uruguay | 2.070 |

Source: Solís-Soberón (1997)

1996). In fact, there are mutual funds that charge 1 percent or less. Second, unlike in the case of mutual funds, in Mexico (as in Chile) membership in an AFORE is compulsory. Thus, even if the transaction costs were the same, comparing the two would be like comparing apples and oranges.

The most pertinent question for this discussion concerns how the transaction costs under the new system compare with costs under the old pay-as-you-go system. Under the old system, administrative costs (the analog of transaction costs) as a percentage of total social security expenditures hovered around 17 percent in the 1980s for Mexico (International Labor Organization 1991). As mentioned above, on average the current transaction cost is around 24 percent, implying a cost increase of roughly 7 percent.

It appears that, at least for the time being, AFOREs are acting as bookkeeping entities with few true portfolio manager functions. As mentioned above, only a very small fraction of Mexican private debt meets CONSAR's eligibility requirements. As is clear in Table 11, the 35 percent ceiling that an AFORE is allowed to hold in the form of private debt is far from binding. In this context, it might make sense for the government to solicit bids and choose the single bookkeeper charging the lowest management fee. While a monopoly in the private management of widely diversified pension portfolios is not desirable, a monopoly in bookkeeping may very well be, as Bolivia's experience demonstrates. Bolivia awarded a monopoly right to an international consortium to manage its pension system.[12] The commission charges are 0.5 percent of average salary (10.5 percent of salary for retirement and 2 percent for life insurance).[13] In contrast, under the Mexican system, the average charges are 1.6 percent of salary with only 6.5 percent going into the retirement fund. In addition, in Bolivia the disability and death insurance payment is equal to 4 percent of salary.

A potential justification for leaving bookkeeping and portfolio management together in the AFOREs could be that the AFOREs will be active when CONSAR relaxes its portfolio restrictions. The problem with this justification is that separating bookkeepers and fund managers may be more efficient even under these conditions. The best bookkeepers may not be the best managers of a wider-spectrum portfolio.

**The Replacement Rate.** It is also important to recognize that because of differences in how the replacement rate is determined—under the new system it will be market-determined—contrasting alternative replacement rate projections under the new system against the replacement rate under the old system will be challenging. The ultimate concern, of course, is whether the new replacement rate will represent an improvement for future retirees. A complete answer requires a sophisticated analysis. The replacement rate under the new regime is the endogenous result of changes in factors such as wage levels (because the value of the social quota depends on the level of wages; that is, a low-wage earner gets a relatively high social quota), real interest rates, mortality rates (because mortality rates are going to change over the next sixty years), the discount rate at which the whole life annuity is calculated, and commissions (for example, some commissions depend on the inflation rate, as for one company that does not charge if there are no gains in real terms).

Table 13 illustrates the sensitivity of the replacement rate to alternative economic scenarios. It reports the resulting replacement rates, other things being equal, under alternative real rates of return of AFOREs' portfolios. These simulations are subject to criticism on a number of grounds. Among other things, they take a flat lifetime wage profile, assume a flat interest rate profile with zero inflation rates, take mortality projections from United Nations (1998), and assume that commissions stay put. However, the simulations suffice to focus attention on the sensitivity of the replacement rate to the relevant economic variables.

Table 13 also has calculations of replacement rates under different scenarios when three elements are altered: (1) The replacement rate is calculated with different real rates of return. However, the effects of inflation are not taken into account, and since earnings of managed funds depend on the nominal interest rate, inflation would have an impact. (2) For each panel of the table, the wage rate varies. Wage rates are expressed as multiples of the minimum wage. So calculations are made with one, two, three, four, five, six, and ten multiples of minimum wages. (3) The number of years in the labor force also varies (from

**TABLE 13 Replacement-Rate Calculations for Whole Life Annuity Starting at Age 65[a]**

| Time (Years) | Salary in Times Minimum Wage | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 10 |
| Real Rate of Return, 3 Percent; Real Wage Growth Rate, 0 Percent | | | | | | | |
| 5 | 3.53 | 2.63 | 2.32 | 2.17 | 2.08 | 2.02 | 1.90 |
| 10 | 7.41 | 5.51 | 4.87 | 4.56 | 4.37 | 4.24 | 3.99 |
| 15 | 11.83 | 8.79 | 7.78 | 7.28 | 6.97 | 6.77 | 6.37 |
| 20 | 16.85 | 12.53 | 11.09 | 10.37 | 9.94 | 9.65 | 9.07 |
| 25 | 22.54 | 16.76 | 14.84 | 13.87 | 13.30 | 12.91 | 12.14 |
| 30 | 29.01 | 21.57 | 19.09 | 17.85 | 17.11 | 16.61 | 15.62 |
| 35 | 36.35 | 27.03 | 23.92 | 22.37 | 21.44 | 20.82 | 19.57 |
| 40 | 44.69 | 33.23 | 29.41 | 27.50 | 26.36 | 25.59 | 24.06 |
| 45 | 54.17 | 40.28 | 35.65 | 33.33 | 31.94 | 31.02 | 29.16 |
| Real Rate of Return, 10 Percent; Real Wage Growth Rate, 0 Percent | | | | | | | |
| 5 | 4.20 | 3.12 | 2.76 | 2.58 | 2.47 | 2.40 | 2.26 |
| 10 | 10.55 | 7.84 | 6.94 | 6.49 | 6.22 | 6.04 | 5.67 |
| 15 | 20.51 | 15.24 | 13.48 | 12.61 | 12.08 | 11.73 | 11.03 |
| 20 | 36.09 | 26.82 | 23.73 | 22.18 | 21.25 | 20.64 | 19.40 |
| 25 | 60.49 | 44.94 | 39.75 | 37.16 | 35.61 | 34.57 | 32.50 |
| 30 | 98.74 | 73.33 | 64.87 | 60.63 | 58.09 | 56.40 | 53.01 |
| 35 | 158.78 | 117.89 | 104.27 | 97.45 | 93.36 | 90.64 | 85.19 |
| 40 | 253.13 | 187.90 | 166.16 | 155.28 | 148.76 | 144.41 | 135.72 |
| 45 | 401.56 | 298.00 | 263.48 | 246.22 | 235.86 | 228.96 | 215.15 |

[a] Calculated with a flat lifetime wage profile and no consideration of inflation.

Source: Sinha (1999b)

five to forty-five years). Note that the wage profile does not vary; the wage rate for every year in the labor force is assumed to be the same.

Consider, for example, the first entry of the top panel of the table—3.53 percent. A person earning the minimum wage for five years and retiring at the age of 65 will get 3.53 percent of his or her wage replaced if he or she earns one minimum wage. Each entry has two other elements built into it. One is the assumption about the real interest rate that an annuity would earn. The other is the (conditional) mortality rate of the population (after retiring at 65). The 3.53 percent results from calculating the replacement rate that would be the average of the rates obtained under each of the seventeen AFOREs. Therefore, each calculation explicitly takes into account management or commission fees.

For understanding the significance of the different replacement rates, it is helpful to compare these replacement rates to the U.S. average. For instance, in the United States in 1998 the average wage was around $28,000. The retirement benefit after forty-three years of service was $11,256. This amounts, roughly, to a 40 percent (11,256/28,000) replacement rate. In Mexico, the average salary is slightly more than three times minimum salary. In U.S. dollars, this amount is around $10 a day. Thus, under the scenario with 3 percent real rate of return, to get a replacement rate of 40 percent the average worker has to work for more than forty-five years. Under the assumption of a 10 percent real rate of return, the 40 percent replacement rate can be achieved in twenty-five years.

12. The consortium consists of Banco Bilbao Vizcaya S.A. and Invesco-Argentaria.
13. See von Gersdorff (1997) for a summary of the Bolivian system.

How does the new system compare with the old in terms of replacement rates? For a person earning one minimum salary, under the old system a 100 percent replacement could be had in ten years (Table 2). Under the new system, if the real rate of return is 3 percent, a person with one minimum salary after forty-five years of service would be able to achieve only a 54.17 percent replacement rate. With a real rate of return of 10 percent, the same person after thirty years of service will get 98.74 percent of the salary replaced. Only if a worker stayed in the labor force for at least thirty-five years and the rate of return was high (10 percent or more) would the retirement benefits under the new system be higher (for almost all wage levels) than those under the old system. One other important observation is that for low-income workers the replacement rate is always higher than for high-income workers. The reason is that as a percentage of income the social quota is financially more important for low-income workers.

## What Are the Net Gains from Switching across Pay-As-You-Go Pension Systems in Mexico?

The last section identified some of the potential gains and costs of the reform. However, what needs to be established are the net gains or overall effects on well-being for current and transitional retirees as well as for future generations of Mexican citizens. Economic outcomes are the result of complex simultaneous interactions among different economic variables in both the short and long runs. To say, for example, that the government contributes a "cuota social" to the retirement fund of a worker (as is currently the case under the reform) is to say that the government commits to borrowing or taxing in the future (from either the same or future generation of workers) to meet the obligation of an accounting entry. Thus, contrasting the well-being of Mexican citizens under the alternative pay-as-you-go systems would require a sophisticated analysis. At a minimum, the analysis should recognize the intertemporal nature of individuals' decision making, that individuals' expectations are based on all available information, including government policies, and that economic variables interact with one another. Thus, to determine the consequences of a reform, one must consider saving decisions, taxing, and government debt policies simultaneously— a general equilibrium framework.

The following example illustrates how such an approach could make a difference. De Nardi, Imrohoroglu, and Sargent (1999) look at the effects of projected U.S. demographics on its current pay-as-you-go system. They use projected increases in the dependency ratio and analyze the economic consequences of several alternative fiscal adjustment packages. One of their experiments consists of leaving the social security system unfunded (perhaps an analog of the Mexican case). They conclude that back-of-the-envelope accounting calculations made outside a general equilibrium framework differ significantly from those obtained in a general equilibrium context. One may therefore wonder about the accuracy of the projected actuarial imbalances—discussed earlier—arising from sticking to the old Mexican pension system.

Another finding in De Nardi, Imrohoroglu, and Sargent is that even when a country sticks to a pay-as-you-go system, "reducing retirement benefits through taxation of benefits and consumption or through postponing the retirement eligibility age results in a significant reduction of the fiscal adjustment necessary to cope with the aging of the population" (1999, 578). As discussed above, under the new Mexican system, there is a new minimum twenty-five year contribution required to qualify for any benefits. This regulation amounts to a reduction of retirement benefits. Also, as discussed above, the Mexican reform may not represent a departure from its pay-as-you-go system. Thus, just as in the comparisons performed in De Nardi, Imrohoroglu, and Sargent, the relevant comparison in Mexico may be across two pay-as-you-go regimes rather than between a pay-as-you-go and a fully funded system. Because that is the case, the relevant analysis would, for example, have to contrast the net benefits of maintaining a pay-as-you-go system while changing the minimum contribution requirement from ten to twenty-five years and introducing the cuota social and other features of the new system reviewed above. It would not be surprising if, just as in De Nardi, Imrohoroglu, and Sargent, a reduction of retirement benefits resulted in a significant reduction of the adjustments necessary to cope with the aging of the population in Mexico. However, De Nardi, Imrohoroglu, and Sargent's findings are specific to the structure and parameters of the U.S. economy, and it seems essential to perform the same experiment for the specific parameters of the Mexican economy. To date, there has not been a general equilibrium analysis of the net benefits of modifying the pay-as-you-go system in Mexico. Thus the answer to the question of what the net gains of a modified pay-as-you-go system might be is that nobody knows.

## What Are the Net Gains from Switching to a Fully Funded Pension System?

As discussed above as well as more thoroughly in Espinosa-Vega and Russell (1999), the theoretical literature suggests that, with

# Comparison of Mexico's Pay-As-You-Go and Reformed Old-Age Security Systems

| Area | Pay-As-You-Go | Reformed |
|---|---|---|
| Institutional responsibilities: | | |
| Old age and severance (RCV) | IMSS | New entrant picks AFORES or IMSS retirement (transition generation only) |
| Disability and life insurance (IV) | IMSS | IMSS |
| Contributions (percent of wage)[a]: | | |
| Contribution by employer and employee | 10.075 | 10.075 |
| Government contribution | 0.425 | 2.425 |
| Eligibility requirements: | | |
| Old age | 500 weeks' (10 years') contribution; 65 years old | 25 years' contribution; 65 years old |
| Severance | 500 weeks' contribution; 60 years old | 25 years' contribution; 60 years old |
| Old age: Withdrawals[b] | | Gradual withdrawals from individual account in AFORE,[c] or annuity bought from an insurance company |
| Minimum pension guarantee (MPG) | Equivalent to one Mexico City minimum-wage level indexed to actual minimum wage | Equivalent to one Mexico City minimum wage on 7/1/97 indexed to the CPI[d] |

[a] Under IVCM, contributions could not exceed ten times the minimum wage, and under the new system the limit is twenty-five times. The column listing the after-reform structure includes Life and Disability Assurance.

[b] Lump withdrawal at retirement permitted only for balances in excess of 130 percent of the cost of an annuity equal to the minimum pension guarantee (MPG).

[c] Only gradual withdrawals are allowed in order to reduce the risk that recipients will outlive their accumulated balances.

[d] Currently average wage for IMSS affiliates is 2.6 minimum wages; thus MPG is approximately 38 percent of average wage.

Sources: Grandolini and Cerda (1998) and Sales-Sarrapy, Solís-Soberón, and Villagómez-Amezcua (1998)

some qualifications, moving to a fully funded system or abolishing a pension system altogether should improve the well-being of society at large. The qualifications include a country having a dynamically efficient economy and the absence of any significant imperfections in capital and labor markets. A number of students of pension systems agree with Abel and others (1989) that the U.S. economy is dynamically efficient but acknowledge that the United States may experience some potentially significant market failures. Huang, Imrohoroglu, and Sargent (1997) start by explicitly incorporating what is generally considered a standard market failure. In their model, there are uninsurable uncertainties about lifetimes and labor income. This type of market failure has indeed been used to justify pay-as-you-go social security systems.

Huang, Imrohoroglu, and Sargent perform two experiments similar to the actions some analysts claim Mexico has started to take. In the first, the government eliminates its pay-as-you-go system, issuing a large quantity of bonds to buy out the transitional generation of retirees. At the same time, the government raises labor income taxes for the next forty years in order to pay back this debt (this tax represents the transition cost). The second experiment is motivated by some recent proposals in the United States that call for investing the current social security surpluses in the stock market. In their second experiment, the government raises labor income taxes. And the proceeds are used to acquire equity. Future retirement benefits are then paid out of the returns to this equity. The advantage of adopting this government-sanctioned fully funded scheme is that it allows the government to redistribute benefits to those citizens who because of, for example, extended layoff periods were unable to accumulate enough savings for their retirement or those citizens who outlived their retirement benefits. They show that in the long run, the second experiment provides the larger net gain. While the study was performed for the U.S. economy parameters (and, as the authors note, there are a number of directions in which it could be improved), its importance for Mexico consists of recognizing that the size of the net gains is itself a function of key features of the economy, including the specific type of market failures in place (such as the large size of the informal sector in Mexico). For example, a common theme throughout most analyses of Mexico's reform is the expected migration from the informal to the formal sector of the economy. In order to understand the impact of the reform on the informal-sector workers' decisions to migrate, one has to endogenize informality. Stated differently, there is no clear idea of the sensitivity of workers' decisions about where to work and the role of changes in payroll taxes associated with alternative pension schemes in those decisions. Neither is it clear how workers' decisions about migrating in turn affect the ultimate impact of government policies in the economy.

At the same time, it may be of significant relevance where the revenues needed to pay for the transition come from. The macroeconomic implications of either taxing wage income or issuing government debt to finance the transition from a pay-as-you-go to a fully funded system may be quite different from each other. Also, it is important to specify the timing and the type of tax to be used to finance the deficit that arises from financing the reform; otherwise, the analysis will be at best incomplete.

Serrano (1999b) attempts to estimate the net benefits of switching to a funded system in Mexico in the context of a general equilibrium analysis.[14] He finds that the gains from doing so significantly outweigh the present value of the transition costs, which in his worst-case scenario represent 59.3 percent of 1997 GDP. Should CONSAR view Serrano's work as their endorsement to press ahead with a reform to a fully funded pension in Mexico? In an effort to answer this question, it is important to outline Serrano's study. As in Huang, Imrohoroglu, and Sargent (1997), Serrano incorporates a market failure in his analysis. The market failure consists of a minimum-denomination restriction in the formal financial intermediation sector under the pay-as-you-go system. That is, he assumes that when such a pension system is in operation, poor savers are unable to participate in the formal banking sector. In his analysis, once a fully funded system is adopted, these poor savers will have access to the formal financial system. In his words, "Our thesis is that many poor workers will obtain access to the formal financial system through the privatized social security system. . . . The introduction of an obligatory FF [fully funded] system may give these people access to the financial system (1999b, 3)." In other words, adopting a fully funded system would have the same benefits, in his analysis, as eliminating the minimum-denomination restriction in the formal banking sector. The question, of course, would be why the poor savers did not have access to the formal financial system in the first place. The answer may be that the low level of savings by the poor was insufficient to justify the necessary maintenance or transaction costs incurred by financial intermediaries when opening a new account. If this is the case, what leads Serrano to believe that things would be different under a fully funded system? The government could, of course, subsidize these transaction

costs, but it could do that regardless of the pension system in place. In addition, the overall impact of such subsidies would have to be carefully analyzed.

It is also true that a large number of poor savers are part of the informal labor market. A government-sanctioned fully funded system is relevant to them only if they migrate to the formal labor market. Serrano would have to explain why such a pension system would lead to migration to the formal sector. Following the steps of Serrano's analysis, one cannot disentangle where the gains come from. Do they come from eliminating the market failure, or do they come from switching from a pay-as-you-go to a fully funded system? In short, it seems best to think of Serrano's work as a work in progress.

Another important point to note is that Serrano's estimate of the transition cost (59.3 percent of 1997 GDP) serves as a reminder that even when there are net gains from switching to a funded system the transition imposes a hefty cost on any society. Agreeing to when and how to pay for it would be the subject of difficult political discourse. It is no coincidence that while there are other estimates of the transition costs, there is no mention (other than Serrano's) about when and how such costs would be taken care of.

In addition to Serrano (1999b), other articles estimate the transition cost. By the authors' own admission, these estimates are far from perfect; for example, Sales-Sarrapy, Solís-Soberón, and Villagómez-Amezcua acknowledge that their model "is a partial equilibrium framework that treats relevant macroeconomic variables as given" (1998, 158). Partial equilibrium estimates may drastically err in either direction, as discussed above and shown by De Nardi, Imrohoroglu, and Sargent (1999). For completeness, this discussion includes transition costs estimates by Sales-Sarrapy, Solís-Soberón, and Villagómez-Amezcua (1998) and Grandolini and Cerda (1998) with a reminder that these costs are only part of the information needed for estimating the net gains of the reform.

Sales-Sarrapy, Solís-Soberón, and Villagómez-Amezcua (1998) and Grandolini and Cerda (1998) acknowledge that moving away from the old pay-as-you-go system will impose on the Mexican economy two types of costs with certainty and another two potential costs. First, the government has to pay the so-called social quota to every participant, and that payout affects the government deficit. As can be seen in the box, this amount on average equals an additional 2 percent of wages. Second, since all contributions of private-sector workers went to AFOREs starting September 1, 1997, one has to consider the resources necessary to provide payments to pensioners existing prior to that date.

The first group of retirees under the new system will emerge in about twenty-five years. What exactly will they get? There are two possible scenarios. In the first, if the funds in the individual's account at retirement do not exceed an annual income stream equivalent to one minimum wage for the actuarial remainder of his life, then the government will guarantee benefits equivalent to one minimum wage for the duration. This approach may affect the government deficit. In the second scenario, if the individual's account exceeds an annual income stream equivalent to one minimum wage for the actuarial remainder of the worker's life, he or she can choose between withdrawing funds on a monthly basis or purchasing a lifetime private contingent annuity. Under this scenario there is no impact on the deficit. Transition workers have a pension-guaranteed switch option. At retirement, they will be able to choose between their benefits under the old or the new system. Their choices may have an impact on the government deficit.

Grandolini and Cerda (1998, table 11, 24) report the fiscal cost of the transition as calculated by the Ministry of Finance (Secretaría de Hacienda y Crédito Público, or SHCP). The present value of the total cost of the transition from 1997 to 2024, as a fraction of 1994 GDP, is 17.76 percent. Sales-Sarrapy, Solís-Soberón, and Villagómez-Amezcua (1998) compute the fiscal deficit arising from the compensation to current pensioners and transition workers from 1997 to 2047 under the new social security scheme. For this period, they estimate that the total cost of the transition would be 82.6 percent of GDP.

After 2047 (the year most transitional workers cease to exist) the only social security cost for the government would be the social quota. If, as estimated in Grandolini and Cerda (1998), this cost were less than 0.1 percent of GDP after 2025, the total social quota cost from 2047 to 2058 would be about 1.1 percent of GDP. Based on the computations presented earlier, the highest estimated cost of the transition from 1997 to 2047 would be 82.6 percent of GDP. Thus, in the worst-case scenario the cost of the transition from 1997 to 2058 would be roughly 84 percent of GDP (1.1 percent plus 82.6 percent).

The message from these computations seems to be that given the fiscal cost savings—even for the worst-case scenario, the reform represents a

14. As of this writing, Serrano's is apparently the only study of its kind.

44 percent savings over the estimated 141 percent of GDP cost of holding on to the status quo—the country should press ahead with the transition to a funded system.

But is this interpretation of these estimates correct? First, as mentioned earlier, detailed attention needs to be given to the different assumptions regarding the type and time of taxes necessary to pay for the transition cost as well as other assumptions about the relevant market imperfections and redistributional considerations. Second, as the discussion above implies, it may be that at best the studies represent estimates of the cost of changing between types of pay-as-you-go systems and not of transitioning to a fully funded system. Since both the estimates of the cost of maintaining the status quo and transitioning to a funded system are partial equilibrium estimates, it is hard to place confidence in them. Finally, in order to evaluate the desirability of either changing the form of the pay-as-you-go system or moving to a fully funded system, it is important to emphasize that what matters is the net benefit of the change in pension system. General equilibrium analyses modeling Mexican unique features are essential if comparative policy analysis is going to be meaningful. Informality in the labor market of the economy and emerging financial markets are two examples of such features that must be accounted for. In closing, on the basis of the information reviewed here, it seems safe to say that nobody really knows what the net gains from switching to a fully funded system might be.

## Conclusion

This article describes some of the factors that if left unchanged very likely would have led to an actuarial imbalance of the Mexican pension system in the near term. This article contends that after reviewing the key aspects of the new pension system, one cannot tell whether the government intends to switch to a fully funded social security system. An interesting question is whether the Mexican public believes the government will carry out such a switch. If the public believes a genuine transition will begin in the future, then they will expect interest rates to be lower in the future: stated differently, the public will begin to view current interest rates as above their long-run levels. This situation should cause them to increase their saving to take advantage of the temporarily high level of interest rates. As they do, the level of private saving should begin to rise and the level of market interest rates should begin to fall.

Unfortunately, the more gradual the public expects the transition to a fully funded system to be, the smaller the expectational effect on saving and interest rates is likely to be. Since Mexico's current financial and political situation would seem to favor a very gradual transition, the effect might not be large enough to be identified easily. It might take many years, or even a generation or more, for the effect on saving and interest rates to be noticeable.

If the conjecture in this article and in Espinosa-Vega and Russell (1999) is correct, the current pension reform is another pay-as-you-go system. It would be interesting, then, to try to estimate the net gains from the switch from the old pay-as-you-go system to a new version of it. Unfortunately, to date there are no solid studies with such estimates. More research is needed to better assess how the impending demographic changes and the changes in the eligibility criteria, the replacement rate, and other aspects of the new system will affect the country's overall economic welfare.

The commitment to switch to a fully funded system is not trivial. It requires decisions about how and when to pay for hefty transition costs. Unfortunately, while there are a number of studies about the effects of switching to a fully funded pension system for the United States, there is little solid information for Mexico. Mexico is in dire need of further research to guide it through its decision on whether and how to switch to a fully funded pension system.

It must be said that Mexico is not alone in this respect. Mexico's experience should be viewed as an illustration of the difficulties in assesssing the net benefits of a pension reform.

## REFERENCES


Abel, Andrew B., N. Gregory Mankiw, Lawrence H. Summers, and Richard J. Zeckhauser. 1989. "Assessing Dynamic Efficiency: Theory and Evidence." *Review of Economic Studies* 56:1–20.

Banco de Mexico. 1996, 1997. "The Mexican Economy." <http://www.banxico.org.mx/gPublicaciones/FSPublicaciones.html>.

CONSAR. 1997. "Circular CONSAR 15-1." *Diario Oficial de la Federacion*, June 30.

———. 1999. <http://www.consar.gob.mx>.

Cooley, Thomas F., and Jorge Soares. 1999. "Social Security Based on Reputation." *Journal of Political Economy* 107 (February): 135–60.

De Nardi, Mariacristina, Selahattin Imrohoroglu, and Thomas J. Sargent. 1999. "Projected U.S. Demographics and Social Security." *Review of Economic Dynamics* 2 (July 1999): 575–615.

Diamond, Peter A. 1965. "Government Debt in a Neoclassical Growth Model." *American Economic Review* 55:1126–50.

———. 1998. "The Economics of Social Security Reform." National Bureau of Economic Research Working Paper no. 6719, September.

Espinosa-Vega, Marco A., and Steven Russell. 1999. "Fully Funded Social Security: Now You See It, Now You Don't?" Federal Reserve Bank of Atlanta *Economic Review* 84 (Fourth Quarter): 16–25.

Feldstein, Martin S. 1974. "Social Security, Induced Retirement, and Aggregate Capital Accumulation." *Journal of Political Economy* 82 (September/October): 905–26.

Grandolini, Gloria, and Luis Cerda. 1998. "The 1997 Pension Reform in Mexico." World Bank Policy Research Working Paper no. 1933, June.

Huang, He, Selahattin Imrohoroglu, and Thomas J. Sargent. 1997. "Two Computations to Fund Social Security." *Macroeconomic Dynamics* 1, no. 1:7–44.

Imrohoroglu, Ayse, Selahattin Imrohoroglu, and Douglas Joines. 1995. "A Life Cycle Analysis of Social Security." *Economic Theory* 6, 83–114.

IMSS (Instituto Mexicano del Seguro Social). 1997. *La Seguridad Social ante el Futuro*. Mexico.

International Labor Organization. 1991. "The Cost of Social Security." Geneva, March.

Judisman, Clara. 1997. "La Informalidad en Mexico: Características y Tendencias." Secretaria del Trabajo. Unpublished document.

Kotlikoff, Laurence J. 1996. "Privatization of Social Security: How It Works and Why It Matters." National Bureau of Economic Research Working Paper no. 5330, October.

Mitchell, Olivia S. 1996. "Administrative Costs in Public and Private Retirement Systems." National Bureau of Economic Research Working Paper no. 5734, August.

Orszag, Peter R., and Joseph E. Stiglitz. 1999. "Rethinking Pension Reform: Ten Myths about Social Security Systems." World Bank Conference on New Ideas about Old-Age Security. September 14–15, 1999. Washington, D.C.: World Bank.

Queisser, Monika. 1998. "The Second Generation Pension Reforms in Latin America." Development Centre Studies, Organisation for Economic Cooperation and Development, Paris.

Rodriguez, L. Jacobo. 1999. "In Praise and Criticism of Mexico's Pension Reform." *Cato Institute Policy Analysis,* no. 340, April 14.

Sales-Sarrapy, Carlos, Fernando Solís-Soberón, and Alejandro Villagómez-Amezcua. 1998. "Pension System Reform: The Mexican Case." In *Privatizing Social Security*, edited by Martin Feldstein. Chicago: University of Chicago Press.

Schwarz, Anita M., and Asli Demirguc-Kunt. 1999. "Taking Stock of Pension Reforms around the World." World Bank. Unpublished paper. May.

Serrano, Carlos. 1999a. "Social Security Reform—How Much Will It Cost and Who Will Pay for It: The Mexican Case." World Bank. Unpublished paper.

———. 1999b. "Social Security Reform, Income Distribution, Fiscal Policy, and Capital Accumulation." World Bank. Unpublished paper.

Sinha, Tapen. 1999a. "Lessons from Privatization of Pension Plans." Paper presented at the Canadian Institute of Actuaries special conference on retirement.

———. 1999b. "We Are Not in Kansas Anymore: Risks of Privatizing Pension." Instituto Tecnológico Autónomo de México. Unpublished paper.

———. Forthcoming. *Privatization of Social Security in Latin America*. Norwell, Mass.: Kluwer Academic Publishers.

Sinha, Tapen, Felipe Martinez, and Constanza Barrios-Muñoz. 1999. "Publicly Mandated Privately Managed Pension in Mexico: Simulations with Transactions Cost." Society of Actuaries, *Actuarial Research Clearing House*, no. 1:323–54.

Sinn, Hans-Werner. 2000. "Why a Funded Pension System Is Useful and Why It Is Not Useful." National Bureau of Economic Research Working Paper no. W-7592, March.

Solís Soberón, Fernando. 1997. "Análisis comparativo de las comisiones que cobrarán las AFORES." CONSAR. Unpublished paper.

United Nations. 1998. *Demographic Bulletin*. Santiago, Chile, July.

Van Ginneken, Wouter. 1998. "Social Security for the Informal Sector: Investigating the Feasibility of Pilot Projects in Benin, India, El Salvador, and Tanzania." Issues in Social Protection Discussion Paper No. 5. Social Security Department, International Labor Office, Geneva, Switzerland.

Von Gersdorff, Hermann. 1997. "Pension Reform in Bolivia: Innovative Solutions to Common Problems." World Bank Policy Research Working Paper no. 1832, September.

World Bank. 1994. *Averting Old Age Crisis*. New York: Oxford University Press.

———. 2000. "Understanding and Responding to Poverty." <http://www.worldbank.org/poverty/mission/up1.htm> (April 4).

# Issues in Hedging Options Positions

**SAIKAT NANDI AND DANIEL F. WAGGONER**

*Nandi is a senior economist and Waggoner is an economist in the financial section of the Atlanta Fed's research department.* *They thank Lucy Ackert, Jerry Dwyer, and Ed Maberly for helpful comments.*

MANY FINANCIAL INSTITUTIONS HOLD NONTRIVIAL AMOUNTS OF DERIVATIVE SECURITIES IN THEIR PORTFOLIOS, AND FREQUENTLY THESE SECURITIES NEED TO BE HEDGED FOR EXTENDED PERIODS OF TIME. OFTEN THE RISK FROM A CHANGE IN VALUE OF A DERIVATIVE SECURITY, ONE WHOSE VALUE DEPENDS ON THE VALUE OF AN UNDERLYING asset—for example, an option—is hedged by transacting in the underlying securities of the option. Failure to hedge properly can expose an institution to sudden swings in the values of derivatives resulting from large unanticipated changes in the levels or volatilities of the underlying assets. Understanding the basic techniques employed for hedging derivative securities and the advantages and pitfalls of these techniques is therefore of crucial importance to many, including regulators who supervise the financial institutions.

For options, the popular valuation models developed by Black and Scholes (1973) and Merton (1973) indicate that if a certain portfolio is formed consisting of a risky asset, such as a stock, and a call option on that asset (see the glossary for a definition of terms), then the return of the resulting portfolio will be approximately equal to the return on a risk-free asset, at least over short periods of time.[1] This type of portfolio is often called a hedge/replicating portfolio. By properly rebalancing the positions in the underlying asset and the option, the return on the hedge portfolio can be made to approximate the return of the risk-free asset over longer periods of time. This approach is often referred to as dynamic hedging. However, forming a hedge portfolio and then rebalancing it through time is often problematic in the options market. There are two potential sources of errors: The first is that the option valuation model may not be an adequate characterization of the option prices observed in the market. For example, the Black-Scholes-Merton model says that the implied volatility should not depend on the strike price or the maturity of the option.[2] In most options markets, though, the implied volatility of an option does depend on the strike price and time to maturity of the option, a phenomenon that runs contrary to the very framework of the Black-Scholes-Merton model itself. The second potential source of error is that many option valuation models, such as the Black-Scholes-Merton model, are developed under the assumption that investors can trade and hedge continuously through time. However, in practice,

investors can rebalance their portfolios only at discrete intervals of time, and investors incur transaction costs at every rebalancing interval in the form of commissions or bid-ask spreads. Rebalancing too frequently can result in prohibitive transaction costs. On the other hand, choosing not to rebalance may mean that the hedge portfolio is no longer close to being optimal, even if the underlying option valuation model is otherwise adequate.

This article examines some strategies often used to offset limitations in the Black-Scholes-Merton model, describing how the risk of existing positions in options can be hedged by trading in the underlying asset or other options. It shows how certain basic hedge parameters such as "deltas," which are defined and discussed later, are derived given an option pricing model. Subsequently, the discussion notes some of the practical problems that often arise in using the dynamic hedging principles of the Black-Scholes-Merton model and considers how investors and traders try to circumvent some of these problems. Finally, the hedging implications of the simple Black-Scholes-Merton model are tested against certain ad hoc pricing rules that are often used by traders and investors to get around some of the deficiencies of the Black-Scholes-Merton model. The Standard and Poor's (S&P) 500 index options market, one of the most liquid equity options markets, is used to compare the hedging efficacies of various models. This study suggests that ad hoc rules do not always result in better hedges than a very simple and internally consistent implementation of the Black-Scholes-Merton model.

## How Are Option Payoffs Replicated and Deltas Derived?

To hedge an option, or any risky security, one needs to construct a replicating portfolio of other securities, one in which the payoffs of the portfolio exactly match the payoffs of the option. Replicating portfolios can also be used to price options, but this discussion will be limited to their hedging properties. Before considering the hedging aspects of the Black-Scholes-Merton model, a few simple examples will illustrate how such portfolios are constructed.

**One-Period Model.**[3] The first example is a European call option on a stock, assuming that the stock is currently valued at $100.[4] In this example, the option expires in one year and the strike or exercise price is $100, and the annual risk-free interest rate is 5 percent so that borrowing $1 today will mean having to pay back $1.05 one year from now. For simplicity, the assumption is that there are only two possible outcomes when the option expires—the stock price can be either $120 (an up state) or $80 (a down state). Note that the value of the call option will be $20 if the up state occurs and $0 if the down state occurs as shown below (see Chart 1).[5]

**Because of the simplicity and tractability of the Black-Scholes-Merton model for valuing options, the model is widely used by options traders and investors.**

Since there are only two possible states in the future, it is possible to replicate the value of the option in each of these states by forming a portfolio of the stock and a risk-free asset. If $\Delta$ shares of the stock are purchased and $M$ dollars are borrowed at the risk-free rate, the stock portion of the portfolio is worth $120 \times \Delta$ in the up state and $80 \times \Delta$ in the down state while $1.05 \times M$ will have to be paid back in either of the states. Thus, to match the value of the portfolio to the value of the option in the two states, it must be the case that

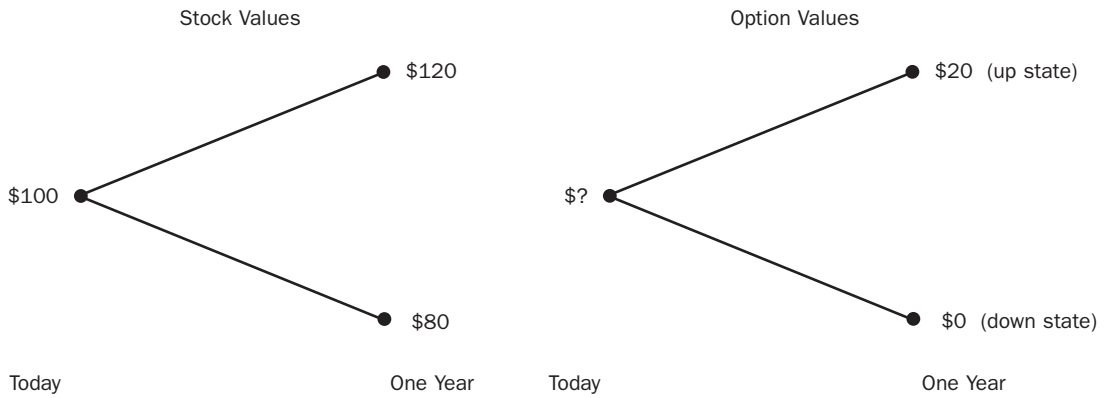
$$120 \times \Delta - 1.05 \times M = 20 \text{ (up state)} \qquad (1)$$

and

$$80 \times \Delta - 1.05 \times M = 0 \text{ (down state)}. \qquad (2)$$

1. For the purposes of this article, the risk-free asset is a money market account that has no risk of default.
2. Implied volatility in the Black-Scholes-Merton model is the level of volatility that equates the model value of the option to the market price of the option.
3. The fact that results reported in this article have been rounded off from actual values may account for small differences when the computations are recreated.
4. The general principle of hedging discussed here applies not only to stock options but also to interest rate options and currency options. Although not discussed here, deltas for American options can be similarly derived for the example shown here. See Cox and Rubinstein (1985) for American options.
5. Note that the risk-free interest rate of 5 percent lies between the return of 20 percent in the up state and –20 percent in the down state. For example, if the interest rate were above 20 percent, then one would never hold the risky asset because its returns are always dominated by the return on the risk-free asset.

**CHART 1 Stock and Option Values in the One-Period Model**




The resulting system of two equations with two unknowns ($\Delta$ and $M$) can be easily solved to get $\Delta = 0.5$, and $M$ is approximately 38.10. Therefore, one would need to buy 0.5 shares of the stock and borrow $38.10 at the risk-free rate in order for the value of the portfolio to be $20 and $0 in the up state and down state, respectively. Equivalently, selling 0.5 shares of the stock and lending $38.10 at the risk-free rate would mean payoffs from that portfolio of –$20 and $0 in the up and down state, respectively, which would completely offset the payoffs from the option in those states.[6] It is also worth noting that the current value of the option must equal the current value of the portfolio, which is $100 \times \Delta - M = \$11.90$.[7] In other words, a call option on the stock is equivalent to a long position in the stock financed by borrowing at the risk-free rate.

The variable $\Delta$ is called the delta of the option. In the previous example, if $C_u$ and $C_d$ denote the values of the call option and $S_u$ and $S_d$ denote the price of the stock in the up and down states, respectively, then it can be verified that $\Delta = (C_u - C_d)/(S_u - S_d)$. The delta of an option reveals how the value of the option is going to change with a change in the stock price. For example, knowing $\Delta$, $C_d$, and the difference between the stock prices in the up and down state makes it possible to know how much the option is going to be worth in the up state—that is, $C_u$ is also known.

**Two-Period Model.** A model in which a year from now there are only two possible states of the world is certainly not realistic, but construction of a multiperiod model can alleviate this problem. As for the one-period model, the example for a two-period model assumes a replicating portfolio for a call option on a stock currently valued at $100 with a strike price of $100 and which expires in a year. However, the year is divided into two six-month periods and the value of the stock can either increase or decrease by 10 percent in each period. The semiannual risk-free interest rate is 2.47 percent, which is equivalent to an annual compounded rate of 5 percent. The states of the world for the stock values are given in Chart 2. Given this structure, how does one form a portfolio of the stock and the risk-free asset to replicate the option? The calculation is similar to the one above except that it is done recursively, starting one period before the option expires and working backward to find the current position.

In the case in which the value of the stock over the first six months increases by 10 percent to $110 (that is, the up state six months from now), the value of the option in the up state is found by forming a replicating portfolio containing $\Delta_u$ shares of the stock financed by borrowing $M_u$ dollars at the risk-free rate. Over the next six months, the value of the stock can either increase another 10 percent to $121 or decline 10 percent to $99, so that the option at expiration will be worth either $21 or $0. Since the replicating portfolio has to match the values of the option, regardless of whether the stock price is $121 or $99, the following two equations must be satisfied:

$$121 \times \Delta_u - 1.0247 \times M_u = 21 \qquad (3)$$

and

$$99 \times \Delta_u - 1.0247 \times M_u = 0. \qquad (4)$$

Solving these equations results in $\Delta_u = 0.9545$ and $M_u = 92.22$. Thus the value of the replicating portfolio is $110 \times \Delta_u - M_u = \$12.78$. If, instead, six months

**CHART 2**
**Stock Values in the Two-Period Model**

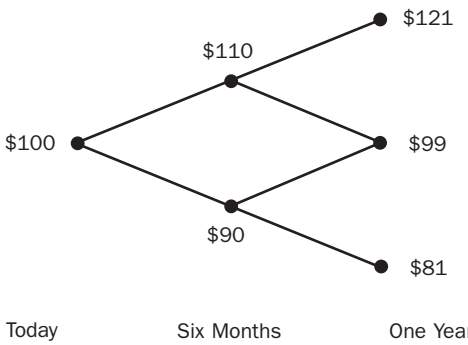
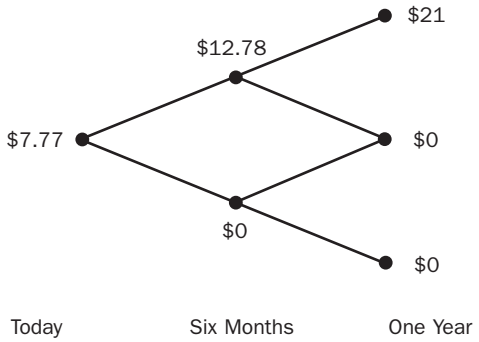



**CHART 3**
**Option Values in the Two-Period Model**




from now the stock declines 10 percent in value, to $90 (the down state), the stock price at the expiration of the option will either be $99 or $81, which is always less than the exercise price. Thus the option is worthless a year from now if the down state is realized six months from now, and consequently the value of the option in the down state is zero. Given the two possible values of the option six months from now, it is now possible to derive the number of shares of the stock that one needs to buy and the amount necessary to borrow to replicate the option payoffs in the up and down states six months from now. Since the option is worth $12.78 and $0 in the up and down states, respectively, it follows that

$$110 \times \Delta - 1.0247 \times M = 12.78, \quad (5)$$

and

$$90 \times \Delta - 1.0247 \times M = 0. \quad (6)$$

Solving the above equations results in $\Delta = 0.6389$ and $M = 56.11$. Thus the value of the option today is $100 \times \Delta - M = \$7.77$. The values of the option are shown graphically in Chart 3.

A feature of this replicating portfolio is that it is always self-financing; once it is set up, no further external cash inflows or outflows are required in the future. For example, if the replicating portfolio is set up by borrowing $56.11 and buying 0.6389 shares of the stock and in six months the up state is realized, the initial portfolio is liquidated. The sale of the 0.6389 shares of stock at $110 per share nets $70.28. Repaying the loan with interest, which amounts to $57.50, leaves $12.78. The new replicating portfolio requires borrowing $92.22. Combining this amount with the proceeds of $12.78 gives $105, which is exactly enough to buy the required 0.9545 ($\Delta_u$) shares of stock at $110 per share. Replicating portfolios always have this property: liquidating the current portfolio nets exactly enough money to form the next portfolio. Thus the portfolio can be set up today, rebalanced at the end of each period with no infusions of external cash, and at expiration should match the payoff of the option, no matter which states of the world occur.

In the replicating portfolio presented above, the option expires either one or two periods from now, but the same principle applies for any number of periods. Given that there are only two possible states over each period, a self-financing replicating portfolio can be formed at each date and state by trading in the stock and a risk-free asset. As the number of periods increases, the individual periods get shorter so that more and more possible states of the world exist at expiration. In the limit, continuums of possible states and periods exist so that the portfolio will have to be continuously rebalanced. The Black-Scholes-Merton model is the limiting case of these models with a limited number of periods.

6. In other words, a long position in one unit of the option can be hedged by holding a short position in 0.5 shares of the stock and lending $38.10 at the risk-free rate: the value of the total position is $0 in both states.
7. If the current value of the option were higher/lower than the value of the replicating portfolio, then an investment strategy could be designed by selling/buying the option and forming the replicating portfolio such that one will always make money at no risk, often called an arbitrage opportunity.

Thus the Black-Scholes-Merton model must assume that investors can trade, or rebalance, continuously through time.[8] Another assumption of the Black-Scholes-Merton model concerns the volatility of the stock returns over each time period. Volatility is related to the up and down movements in the limited-period models. The Black-Scholes-Merton model assumes that the volatility of the stock returns is either constant or varies in such a way that future volatilities can be anticipated on the basis of current information.[9]

Although the continuous trading assumption may seem unrealistic, the Black-Scholes-Merton model nevertheless provides traders and investors with a very convenient formula in which all the input variables but one are observable. The only unobservable input variable is the implied volatility, that is, the average expected volatility of the asset returns until the option expires. A reasonable guess about the expected future volatility is not very difficult, however, because one can estimate the prevalent volatility from the history of asset prices to the present time. From a trader's or investor's perspective, using the Black-Scholes-Merton formula, then, requires only guessing the implied volatility.[10] A more sophisticated option pricing model, in contrast, may require the trader to guess values of model variables more difficult to obtain in real time, such as the speed of mean reversion of volatility and others. In fact, the simplicity of the Black-Scholes-Merton model largely explains its widespread use regardless of some of its glaring biases from a theoretical perspective. Despite the Black-Scholes-Merton model's very convenient pricing formula, it seems to have serious constraints: it does not allow forming a self-financing replicating portfolio with the provision that one can trade only at discrete intervals of time with nonnegligible transaction costs such as commissions or bid-ask spreads.

**Delta Hedging under the Black-Scholes-Merton Model.** Considering a European call option on a nondividend paying stock will illustrate some of the shortcomings of the Black-Scholes-Merton model.[11] This example assumes that the option has a strike price of $100 and expires in 100 days; that the current stock price is $100 and the implied volatility is 15 percent annually; and that the current annual risk-free rate, continuously compounded, is 5 percent. If 100 call options have been written (100 options typically constitute an options contract), a delta-neutral portfolio will have to be formed to hedge exposure to stock price movements. A delta-neutral portfolio is one that is insensitive to small changes in the price of the underlying stock. Using the Black-Scholes-Merton option valuation formula given in Box 1, the value of each option is approximately $3.8375, so that $383.75 is received by selling or writing the option. Since the portfolio should be self-financing, the proceeds from the options are invested in the stock and risk-free asset. Thus $383.75 is invested in a portfolio of $N$ shares of the stock and in $M$ dollars of the risk-free asset.

Let $\Delta$ denote the delta of the option and, in accordance with the formula for $\Delta$ for the Black-Scholes-Merton model given in Box 1, $\Delta = 0.5846$. The delta of the total position (option, stock, and risk-free asset) is a linear combination of the deltas of the options, the stock, and the risk-free asset. The delta of a long (short) position in the option is $\Delta$ ($-\Delta$), the delta of a long (short) position in the stock is 1 ($-1$), and the delta of the risk-free asset is zero. As 100 options have been sold and $N$ shares have been bought, the delta of the portfolio is $-100 \times \Delta + N$.

In order for the portfolio to be delta-neutral, the following equation must be satisfied:

$$-100 \times \Delta + N = 0. \tag{7}$$

Similarly, for the portfolio to be self-financing, it has to be the case that

$$N \times 100 + M = 383.75. \tag{8}$$

In solving the two equations above for $N$ and $M$, $N = \Delta \times 100 = 58.46$ and $M = -5{,}462.25$. Thus 100 options have been sold for a total of $383.75, 58.46 units of the share have been bought, and $5,462.25 has been borrowed at an annual interest rate of 5 percent. The total value of the portfolio is zero when it is formed because the portfolio is self-financing. What happens, though, to the portfolio value on the next trading day for three different levels of the stock prices? Borrowing $5,462.25 has incurred interest charges of approximately $5{,}462.25 \times 0.05/365.0 = \$0.748$. Thus the value of the portfolio on the next day (denoted as $t + 1$) is

$$V(t+1) = 58.46 \times S(t+1) - 100 \times C(t+1) - (5{,}462.25 + 0.748), \tag{9}$$

where $S(t + 1)$ and $C(t + 1)$ denote the values of the stock and the call option, respectively, on the next day. Table 1 gives the value of the option and thereby the value of the delta-neutral portfolio for various values of the stock price, assuming that everything else (including the implied volatility) is the same.

The value of the delta-neutral portfolio is not zero in any of these cases, even though in one the stock price did not change from its initial value of $100. The reason is that the delta has been derived from a

model that assumes continuous trading and thus requires continuous rebalancing for the delta-neutral portfolio to retain its original value. Transactions costs, like broker commissions and margin requirements, would further deteriorate the performance of the delta-neutral portfolio.[12]

**Other Dynamic Hedging Procedures Using the Black-Scholes-Merton Model.** The previous example assumed that the underlying Black-Scholes-Merton model generated the option prices so that the implied volatility was the same on both days. However, in reality the implied volatility is not constant but changes through time in almost all options markets. The following example demonstrates the outcome if the implied volatility changes on the next day, assuming that the implied volatility on the next day ($t + 1$) is 15.5 percent, 15 percent, and 14.5 percent, corresponding to three different stock prices of $99, $100, and $101. The fluctuation of implied volatility suggested here corresponds to stock price, increasing as the stock price goes down and decreasing as it goes up—a feature of many equity and stock index options markets. Table 2 shows the values of the portfolio corresponding to three different levels of stock prices and implied volatilities.

Thus, with a change in the implied volatility of around 0.5 percent (frequently observed in options markets), the hedging performance of the Black-Scholes-Merton model deteriorates quite sharply. The hedge portfolios constructed on the previous day are quite poor primarily because the model's assumption of constant variance is violated. Extensive academic literature documents how implied volatilities in the options market change through time (Rubinstein 1994; Bates 1996; and many others).[13] Further, volatility often varies in ways that cannot always be predicted with current information. How could traders or investors set up hedge portfolios that would account for the random variation in volatilities? One alternative is to derive the hedge portfolio from a more sophisticated (and more complex) option pricing model such as a stochastic volatility model (to be discussed later). However, estimating and implementing such a model can be difficult for an average trader or investor. Practitioners may be better served by finding ways to circumvent the hedging deficiencies of the Black-Scholes-Merton model stemming from implied volatilities that change through time but sticking to the model as much as possible.

One way to get around the problem of time-varying volatility that occurs with the Black-Scholes-Merton model is to form a hedge portfolio that is insensitive to both the changes in the price of the underlying asset and its volatility. The sensitivity of an option price with respect to the volatility is often referred to as vega. In order to hedge against changes in both the asset price and volatility, one can form a portfolio that is delta-neutral as well as

**TABLE 1**
**The Delta-Neutral Portfolio on the Next Day with No Change in Implied Volatility**

| Stock Price | Option Price | Portfolio Value |
|---|---|---|
| $ 99 | $3.26 | –$0.96 |
| $100 | $3.82 | $1.53 |
| $101 | $4.42 | –$0.84 |

**TABLE 2**
**The Delta-Neutral Portfolio on the Next Day When Implied Volatility Changes**

| Stock Price | Implied Volatility (Percent) | Portfolio Value |
|---|---|---|
| $ 99 | 15.5 | –$11.26 |
| $100 | 15.0 | $ 1.50 |
| $101 | 14.5 | $ 9.06 |

8. This replication with continuous trading is possible due to a special property known as the martingale representation property of Brownian motions (see Harrison and Pliska 1981).
9. However, with continuous trading, one can form a self-financing portfolio by trading in the stock and the risk-free asset even if the volatility of the stock is random. All that is needed is that the Brownian motions driving the stock price and the volatility are perfectly correlated (see Heston and Nandi forthcoming).
10. Given the existence of multiple implied volatilities from different options (on the same asset), this task is a little more complicated.
11. If the stock pays dividends, then the present value of the dividends that are to be paid during the life of the option must be subtracted from the current asset price; the resulting asset price is used in the option pricing formula.
12. It is also worth noting that the portfolio is not self-financing on the next day because rebalancing would incur an external cash flow in each of the three states.
13. One can also go to the Web site www.cboe.com/tools/historical/vix1986.txt to see the daily history of the implied volatility index on the Standard and Poor's 100, called the VIX. VIX captures the implied volatilities of certain near-the-money options on the Standard and Poor's 100 index (ticker symbol, OEX).

BOX 1

# Black-Scholes Price and Deltas

The Black-Scholes-Merton formula gives the current value of a European call/put option in terms of (a) $S(t)$, the price of the underlying asset; (b) $K$, the strike or exercise price; (c) $\tau$, the time to maturity of the option; (d) $r(\tau)$, the risk-free rate or the equivalent yield of a zero-coupon bond (that matures at the same time as the option); and (e) $\sigma$, the square root of the average per period (for example, daily) variance of the returns of the underlying asset that will prevail until the option expires.[1] Assuming that the underlying asset does not pay any dividends until the option expires, the call and put values are at time $t$.

$$C(t) = S(t)\,N(d1) - K\exp[-r(\tau)\tau]\,N(d2), \quad \text{(B1)}$$

and

$$P(t) = K\exp[-r(\tau)\tau]\,N(-d2) - S(t)\,N(-d1), \quad \text{(B2)}$$

where $N()$ is the standard normal distribution function and

$$d1 = \{\ln(S/K) + [r(\tau) + 0.5\sigma^2]\tau\}/\sigma\sqrt{\tau} \quad \text{(B3)}$$

and

$$d2 = d1 - \sigma\sqrt{\tau}. \quad \text{(B4)}$$

(The tables for computing the function are found in almost all basic statistics books.) If the underlying asset pays known dividends at discrete dates until the option expires, then the present value of the dividends must be subtracted from the asset price to substitute for $S(t)$ in the above formulas.[2] Of the above-mentioned variables that are required as inputs to the Black-Scholes-Merton formula, only $\sigma$ is not readily observable.

The delta of the option is the partial derivative of the option price with respect to the asset price, that is, $dC/dS$ for call options and $dP/dS$ for put options. An important property of the Black-Scholes-Merton formula is that the option price is homogeneous of degree 1 in the asset price and the strike price. Hence it follows from Euler's theorem on homogeneous functions (see Varian 1984) that the delta of the call option is $N(d1)$ and that of the put option is $N(d1) - 1$.

The vega of a call or put option is $dC/d\sigma$ or $dP/d\sigma$. Hull (1997, 329) gives the formula for vega in terms of the same variables that appear in the valuation formula.

1. Actually the Black-Scholes (1973) model assumes that the risk-free rate is constant. However, Merton (1973) shows that even if interest rates are random, the appropriate interest rate to use in the Black-Scholes formula for a stock option is the yield of a zero-coupon bond that expires at the same time as the option. In that case, the simple Black-Scholes (1973) formula serves as an extremely good approximation because the volatility of interest rates is relatively low compared with the volatility of the underlying stock.
2. The corresponding exact valuation formula for American put options (or call options on dividend paying assets) and deltas are not known explicitly. However, there are good analytical approximations as in Carr (1998), Ju (1998), and, Huang, Subrahmanyam, and Yu (1996).

vega-neutral. The formation of such a portfolio is indeed ad hoc: in fact, it is theoretically inconsistent because under the Black-Scholes-Merton model volatility is constant (or deterministic) and therefore does not need to be hedged. Forming a delta-vega-neutral portfolio would require trading two options, the underlying asset and the risk-free asset.

Adding to the previous example, in which an option contract has been sold (with 100 days to expire) and in which all other variables such as the stock price and the strike price are the same as before, $N2$ units of a second option, $N3$ units of the stock, and $M$ dollars of the risk-free asset are required. The current values of the first and second option are denoted as $C(1)$ and $C(2)$, respectively, whereas the current stock price is denoted as $S(t)$. Since the second option can be chosen freely, an option of the same strike ($100) but a maturity of 150 days is selected. Given these, $C(1)$ = $3.8375 and $C(2)$ = $4.898. The current deltas of the two options are denoted as $\Delta(1)$ and $\Delta(2)$, and the vegas, as vega(1) and vega(2) (see Hull 1997 for the formula for vega).

For the delta of the portfolio to be zero, it is necessary that

$$-100 \times \Delta(1) + N2 \times \Delta(2) + N3 = 0. \quad (10)$$

For the vega of the portfolio to be zero, it is necessary that

$$-100 \times \text{vega}(1) + N2 \times \text{vega}(2) = 0.^{14} \quad (11)$$

For the portfolio to be self-financing, it is necessary that

$$-100 \times C(1) + N2 \times C(2) + N3 \times S(t) - M = 0. \quad (12)$$

Solving the equations in this system of three equations with three unknowns ($N2$, $N3$, and $M$) shows that $N2 = 82.59$, $N3 = 8.64$, and $M = \$884.96$. Thus 82.59 units of the second option and 8.64 units of the stock must be bought, and $884.96 must be borrowed at the risk-free rate. Table 3 shows the value of the delta-vega-neutral portfolio on the next day. The terms $C(1)_{t+1}$ and $C(2)_{t+1}$ denote the prices of the first and second option on the next day.

The delta-vega-neutral hedge portfolio performs much better than a delta-neutral hedge portfolio that uses just one option, especially if the implied volatilities change. The only disadvantage in using this kind of hedge is that the portfolio requires two options, and options markets tend to be less liquid than the market on an underlying asset, such as a stock. On average, options have much higher bid-ask spreads (relative to their transaction prices) than those on an underlying asset such as a stock. Using a second option to hedge the volatility risk therefore could increase transaction costs, especially for a retail investor.

Similar to delta-vega hedging is what is known as delta-gamma hedging. The gamma of an option measures the rate of change of its delta with respect to a change in the price of the underlying asset. The more the delta of the option changes with the asset price, the more a portfolio will have to be rebalanced to remain delta-neutral. The purpose of delta-gamma hedging is to create a portfolio that is both delta-neutral and gamma-neutral. Thus, ceteris paribus, the amount of rebalancing required in a delta-gamma-neutral portfolio would tend to be lower than that in a delta-neutral portfolio over short periods of time, and lower rebalancing could be used to offset higher transactions costs. Constructing a delta-gamma-neutral portfolio also requires two options; the number of units of the second option can be found by solving a similar set of equations to those applied to the delta-vega-neutral portfolio discussed previously. A delta-vega-gamma-neutral portfolio can also be created, but forming such a portfolio requires positions in three options.

**TABLE 3**
**The Delta-Vega-Neutral Portfolio on the Next Day When Implied Volatility Changes**

| Stock Price | Implied Volatility (Percent) | $C(1)_{t+1}$ | $C(2)_{t+1}$ | Portfolio Value |
|---|---|---|---|---|
| $ 99 | 15.5 | $ 3.36 | $ 4.42 | $ –0.30 |
| $100 | 15 | $ 3.81 | $ 4.88 | $ 0.51 |
| $101 | 14.5 | $ 4.32 | $ 5.38 | $ –0.34 |

The hedging problems discussed thus far fall under the rubric of dynamic hedging in that they require a portfolio formed of the underlying asset and a risk-free asset or options that must be rebalanced through time. Since the number of units of the underlying asset and the risk-free asset or other options are derived from an option pricing model, such as the Black-Scholes-Merton model, the formation of the hedge portfolio is prone to model misspecifications; that is, the underlying options valuation model is not consistent with the option prices observed in the market. An alternative to dynamic hedging is static hedging in which a portfolio is formed as of today and requires no further trading in the underlying asset and options.

Let $S$, $K$, $P$, and $C$ denote the underlying asset price, strike price, put price, and call price, respectively. (Note that both the put and call have the same strike price.) If $r$ and $\tau$ denote the risk-free rate and time to expiration, then the put-call parity relationship for European options states that the following has to hold exactly at any given point in time (in the absence of transactions costs):

$$P = C - S + Ke^{-r\tau}. \quad (13)$$

Thus, to replicate the payoff of a put option with the strike price, $K$, and time to maturity, $\tau$, a synthetic portfolio must be constructed containing a call option of the same strike and maturity as that of the put, a short sell of the asset, and a long position on $K$ units of a discount bond (that pays off $1 at maturity) that matures at the same time as the options. Once the synthetic portfolio has been set up for the put option, rebalancing is no longer necessary because the price of the put option is identical to that of the synthetic portfolio if put-call parity is to be preserved. Since the put-call parity relationship is

14. The vega of a portfolio of options is a linear combination of the vegas of the individual options, and the vega of the underlying asset is zero. Vega(1) = 20.41; vega(2) = 24.71; $\Delta(1) = 0.585$; $\Delta(2) = 0.603$.

independent of any option valuation model, static hedging may seem to be the preferable path. However, static hedging is also prone to some of the same drawbacks that occur when options are hedged with options—namely, that options markets are relatively illiquid, and the second option may not be available in the right quantity. For example, in the Standard and Poor's 500 index options, a market maker may have to satisfy huge buy order flows in deep out-of-the-money put options—those with strike prices substantially below the current S&P 500 level—from institutional investors who want to hedge their positions against sharp downturns in the index. However, the volume of deep-in-the-money call options that would be required in the hedge/replicating portfolio (as per put-call parity) is relatively low, and hedging deep-out-of-the-money puts via deep-in-the-money calls may not be readily feasible.

Static hedging has often been advocated as a useful tool for certain types of exotic options known as barrier options.[15] Barrier options tend to have regions of very high gammas; that is, the delta changes very rapidly and thus requires frequent rebalancing in certain regions (for example, if the asset price is close to the barrier). Dynamic hedging may therefore turn out to be quite difficult and costly for barrier options. Nevertheless, liquidity issues concerning static hedging discussed previously also apply to barrier options. A further difficulty is that some options needed as part of the static hedge portfolios for barrier options may not be traded at all, so close substitutes must be chosen. In hedging exotic options such as barrier options, a trade-off between the pros and cons of static and dynamic hedging is thus inevitable.

**Smile, Smirk, and Hedge.** Because of its simplicity (traders have to guess only one unobservable variable—the average expected volatility of the underlying asset over the life of the option) the Black-Scholes-Merton model continues to be very popular with most traders. However, from a theoretical perspective, the model always exhibits certain biases. One very prevalent and widely documented bias is that the implied volatilities in the Black-Scholes-Merton model depend on the strike price and maturity of an option. Chart 4 shows the implied volatilities in the Standard and Poor's 500 index options for call options of different strike prices on December 21, 1995, with twenty-eight and fifty-six days to maturity. The implied volatilities in the Standard and Poor's 500 index options market tend to decrease as the strike price increases; this pattern is sometimes referred to as a volatility smirk. Similarly, in some other options markets, such as the currency options market, the implied volatilities decrease initially as the strike price increases and then increase a little—a U-shaped pattern often referred to as a smile. Chart 4 also makes it apparent that for options of the same strike price, implied volatility differs depending on the maturity of the option. For example, if the strike price is $570, the implied volatility of the option with twenty-eight days to maturity is 18.7 percent whereas the implied volatility of the option with fifty-six days to maturity is 16.7 percent. Such variations in implied volatilities across strike prices and maturities are inconsistent with the basic premise of the Black-Scholes-Merton model, which accommodates only one implied volatility irrespective of strike prices and maturities. Before examining the hedging implications of this bias, it is important to understand what could possibly be causing such a phenomenon for index options.

One possibility for the existence of the smirk pattern in implied volatilities is that the options market expects the Standard and Poor's 500 index to go down with a higher probability than that suggested by the statistical distribution postulated for the returns of the index in the Black-Scholes-Merton model. As a result, the market would put a higher price on an out-of-the-money put than would the Black-Scholes-Merton model. Since option prices (both puts and calls) under Black-Scholes-Merton increase as volatility increases, the implied volatility using the Black-Scholes-Merton model would be higher than it would otherwise be. In fact, if the distribution of the returns of the underlying asset is seen as embedded in a cross section of option prices with different strike prices (see Jackwerth and Rubinstein 1996), the distribution appears to be one in which, given today's index level, the probability of negative returns in the future is higher than the probability of positive returns of equal magnitude. Such distributions are said to be skewed to the left.[16] In contrast, the statistical distribution that drives the returns of an underlying asset under the Black-Scholes-Merton model is Gaussian/normal, which does not involve skewness. In other words, given today's index level, the probability of positive returns is the same as the probability of negative returns of equal magnitude.

Is it possible to get such negatively skewed distributions under alternative assumptions of the statistical process that generates returns? It turns out that allowing for future changes in volatility to be random and allowing volatility to be negatively correlated with the returns of the underlying asset can generate negatively skewed distributions of the returns of the underlying asset.[17] Indeed, option pricing models have been developed in which the

volatility of the underlying asset varies randomly through time and is correlated with the returns of the underlying asset. One class of such models, known as implied binomial tree/deterministic volatility models, was first proposed by Dupire (1994), Derman and Kani (1994), and Rubinstein (1994). In these models the current volatility (sometimes known as local volatility) is a function of the current asset price and time, unlike in the Black-Scholes-Merton model, in which volatility is constant through time.[18] These models are also known as path-independent time-varying volatility models in that the current volatility does not depend on the history or path of the asset price. In another class of models, sometimes known as path-dependent time-varying volatility models, the current volatility is the function of the entire history of asset prices and not just the current asset price.[19]

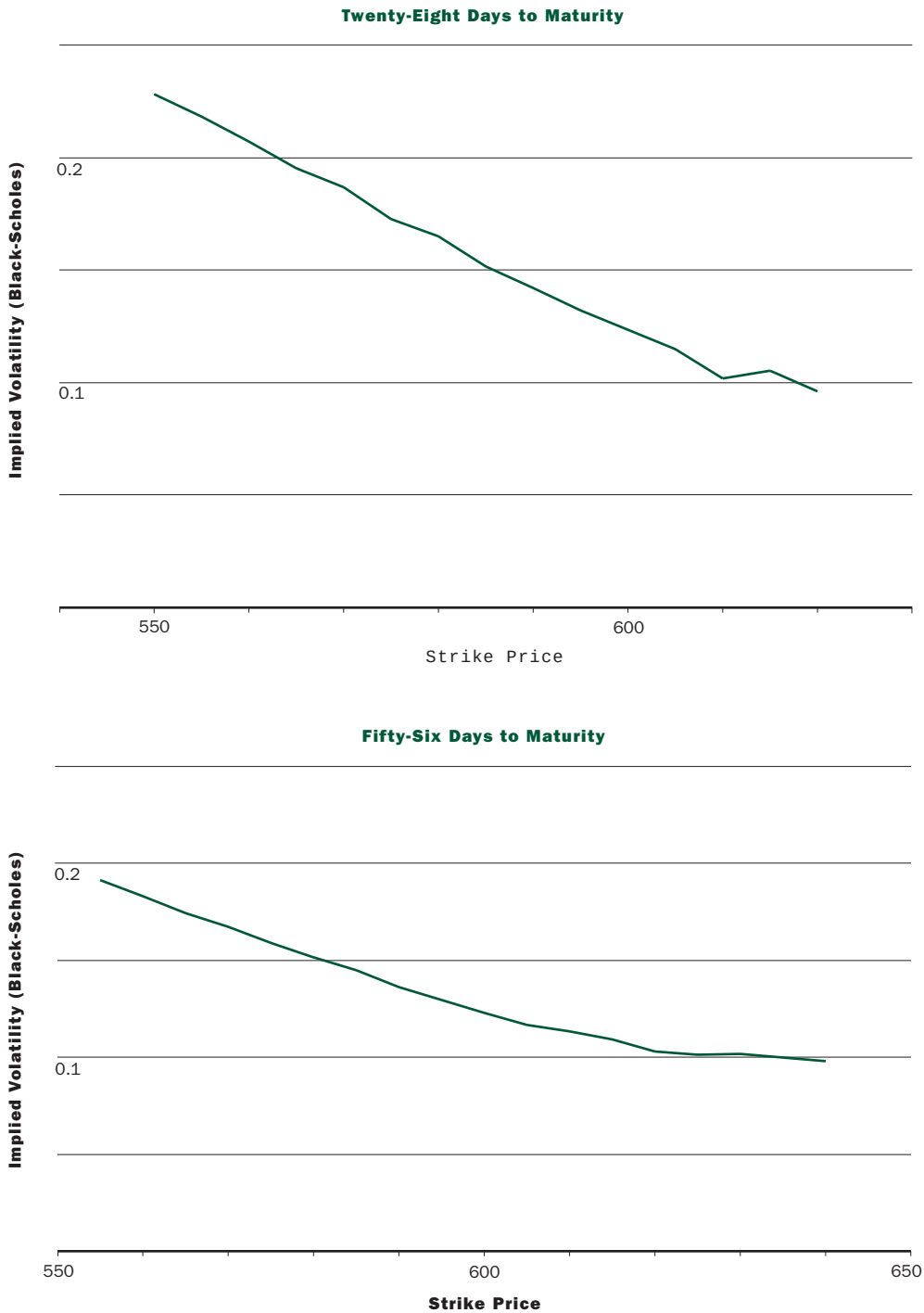
Testing the hedging efficacy of an option valuation model often involves measuring the errors incurred in replicating the option with the prescribed replicating portfolio of the model. In other words, the replicating portfolio is formed today, and at a future time the value of the replicating portfolio is compared with the option price observed in the market as of that time. In empirical tests of path-independent time-varying volatility models, Dumas, Fleming, and Whaley (1998) show that in the Standard and Poor's 500 index options market the replication errors of delta-neutral portfolios of path-independent volatility models are greater than those of the very simple Black-Scholes-Merton model. In fact, in terms of replication errors of delta-neutral portfolios, a very simple implementation of the model also appears to dominate an ad hoc variation of the model that uses a separate implied volatility for each option to fit to the smile/smirk curve. The Black-Scholes-Merton model proves more useful for hedging despite the fact that in terms of predicting option prices (that is, computing option prices out-of-sample) it is dominated by the ad hoc rule and the time-varying path-independent volatility model.

Why is it more useful? As discussed above, the hedge ratio, or the delta, which measures the rate of the change in option price with respect to the change in the price of the underlying asset, is an important consideration. If a replicating/hedge portfolio (from an option pricing model) is formed to replicate the value of the option at the next period, it can be shown that to a large extent the hedging/replication error reflects the difference in the pricing or valuation error between the two periods (see Dumas, Fleming, and Whaley 1998). Though one model, model A for example, may result in a lower pricing error (even out-of-sample) than another model, in order for model A to result in lower hedging errors than model B, it could also often be necessary that the change (across two time periods) in valuation error under model A be less than that under model B. More often than not, however, the differences in the valuation errors (across two time periods) between models turn out not to be very significant for most classes of options (that is, options of different strike prices and maturities). In other words, although the Black-Scholes-Merton model exhibits pricing biases, as long as these biases remain relatively stable through time, its hedging performance can be better than the performance of a more complex model that can account for many of the biases, especially if the more complex model does not adequately characterize the way asset prices evolve over time.

**Hedging with Ad Hoc Models.** How do traders or investors who routinely use the Black-Scholes-Merton model to arrive at hedge ratios/deltas use the model, despite the fact that patterns in implied volatilities across options of different strike prices

15. An example of a barrier option is a down-and-out call option in which a regular call option gets knocked out; that is, it ceases to exist if the asset price hits a certain preset level.
16. The distribution that is skewed is the risk-neutral distribution of asset returns (see Nandi 1998 for risk-neutral probabilities/distributions) and not necessarily the actual distribution of asset returns.
17. Negative correlation implies that lower returns are associated with higher volatility. As a result, the lower or left tail of the distribution spreads out when returns go down, generating negative skewness. This negative correlation is often referred to as the leverage effect (Black 1976; Christie 1982) in equities. One possible explanation for this effect is that as the stock price goes down, the amount of leverage (ratio of debt to equity) goes up, thus making the stock more risky and thereby increasing volatility. An argument against this explanation is that the negative correlation can be observed for stocks of corporations that do not have any debt in their capital structure.
18. Since the future level of the asset price is unknown, the future local volatility is also not known, and, strictly speaking, unlike in the Black-Scholes-Merton model, volatility is not deterministic in these models.
19. See Heston (1993) and Heston and Nandi (forthcoming) for option pricing models with path-dependent volatility models in continuous and discrete time, respectively. These models are sometimes known as continuous time stochastic volatility and discrete-time GARCH models, respectively. Continuous time models are very difficult to implement due to the fact that volatility is unobservable given the history of asset prices.

## CHART 4 Implied Volatilities of Call Options




Note: The chart shows the implied volatilities from Standard and Poor's 500 call options of different strike prices on December 21, 1995. The Standard and Poor's 500 index level was at approximately 610.

BOX 2

# Parameter Estimation

The Black-Scholes-Merton-2 version of the model uses a procedure called nonlinear least squares (NLS) to estimate a single implied volatility across all options each Wednesday. The NLS procedure minimizes the squared errors between the market option prices and model option prices. The difference between the model price (given an implied volatility, $\sigma$) and the observed market price of the option is denoted by $e_i(\sigma)$. As mentioned in Box 1, the midpoint of the bid-ask quote is used for the observed market price of the option. Thus the criterion function minimized at each $t$ (over $\sigma$) is

$$\sum_{i=1}^{N_t} e_i(\sigma)^2,$$

where $N_t$ is the number of sampled bid-ask quotes on day $t$. In essence, this procedure attempts to find a single implied volatility that minimizes the squared pricing errors of the model.

and maturities are inconsistent with the model? As it turns out, such traders or market makers often use certain theoretically ad hoc variations of the basic Black-Scholes-Merton model to circumvent its biases. Such ad hoc variations allow the implied volatilities input to the Black-Scholes-Merton model to differ across strike prices and maturities. Using a separate implied volatility for each option is inconsistent with the basic theoretical underpinning of the Black-Scholes-Merton model, but it is a common practice among traders and market makers in certain options exchanges (Dumas, Fleming, and Whaley 1998). In the course of implementing such ad hoc variations, options traders or investors can be thought of as using the Black-Scholes-Merton model as a translation device to express their opinion on a more complicated distribution of asset returns than the Gaussian distribution that underlies the Black-Scholes-Merton model.

Ad hoc variations of the basic Black-Scholes-Merton model, depending on the way they are designed, may result in prices that better match observed market prices. But do they necessarily result in better hedging performance? Four versions of the Black-Scholes-Merton model that differ from one another in terms of fitting a cross section of option prices (in-sample errors) and also in predicting option prices (out-of-sample errors) will be presented; these examples illustrate that the differences between the models in terms of hedging/replication errors are not as significant as the differences in valuation errors for most options. In fact, if the models are ranked in terms of the replication errors of the delta-neutral portfolios, the ranking could prove different than when the models are ranked in terms of valuation errors.

There are many different ways in which a trader or investor can input a value for volatility in the Black-Scholes-Merton formula for computing the delta of an option. The Black-Scholes-Merton model assumes that the volatility of an asset's returns is constant through time. However, an investor trying to use the model in the real world is not constrained to hold the volatility constant and can periodically estimate volatility from past observations of asset prices. As an alternative to using the historical data, a single implied volatility for all options (of different strikes and maturities every day) can be estimated that minimizes a criterion function involving the squared price differentials between model prices and the observed prices in the market (see Box 2 for details).

This approach results in a single implied volatility for all options every day. On the other hand, implied volatility can be based on observation of a particular option so that a different implied volatility exists for each option. As an alternative to using the exact implied volatility for each option, a procedure that "merely smoothes Black/Scholes implied volatilities across exercise prices and times to expiration" is used by some options market makers at the Chicago Board Options Exchange (CBOE) (Dumas, Fleming, and Whaley 1998). For example, given that the shape of the smirk in implied volatilities resembles a parabola, one can choose the implied volatility to be a function of the strike price and the square of the strike price. However, implied volatilities differ across maturities even for the same strike price. Thus the time to maturity—and possibly the square of the time to maturity—can also be included in the function. The equation below is used in Dumas, Fleming, and Whaley (1998).

$$\sigma(K, \tau) = a_0 + a_1 K + a_2 K^2 + a_3 \tau + a_4 \tau^2 + a_5 K\tau, \quad (14)$$

where $K$ is the strike price and $\tau$ is the time to maturity of the option. Since the implied volatility $\sigma(K,\tau)$ is observable for each $K$ and $\tau$, one can use the above equation as an ordinary least squares (OLS) regression of the implied volatilities on the various right-hand variables to get the coefficients $a_0$, $a_1$, $a_2$, and so on. These coefficients provide an estimated implied volatility for each option.[20] To summarize, one can use the Black-Scholes-Merton model to arrive at the delta in four different ways: (a) compute the delta with volatility estimated from historical prices, (b) compute the delta using a single implied volatility that is common across all options, (c) compute the delta using the exact implied volatility for each option, and (d) compute the delta using an estimated implied volatility for each option that fits to the shape of the smirk across strike prices and time to maturities.

Of the four different versions of the Black-Scholes-Merton discussed above, the two that allow implied volatilities to differ across options of different strike prices and maturities are indeed ad hoc. The other two versions that result in a single implied volatility across all strikes and maturities are much less ad hoc. Implementing the four different versions of the Black-Scholes-Merton model in the Standard and Poor's 500 index options makes it possible to explore the differences in hedging errors produced by these approaches.

The market for Standard and Poor's 500 index options is the second most active index options market in the United States, and in terms of open interest in options it is the largest. It is also one of the most liquid options markets.[21] These models test data for the time period from January 5, 1994, to October 19, 1994.[22] Box 3 gives a detailed description of the options data used for the empirical tests. The replicating/hedge portfolios are formed on day $t$ from the first bid-ask quote in that option after 2:30 P.M. (central standard time). The portfolio is liquidated on one of the following days—$t + 1$, $t + 3$, or $t + 5$.[23] The hedging error for each version of the Black-Scholes-Merton model is the difference between the value of the replicating portfolio and the option price (measured as the midpoint of the bid-ask prices) at the time of the liquidation.

The first panel of Table 4 shows the mean absolute hedging errors (for the whole sample and across all options) of the four versions of the Black-Scholes-Merton (BSM) model.[24] Black-Scholes-Merton-1 is the version of the model in which volatility is computed from the last sixty days of closing Standard and Poor's 500 index levels. Black-Scholes-Merton-2 is the version of the model in which a single implied volatility is estimated for all options each day. Ad hoc-1 is the ad hoc version of the Black-Scholes-Merton model in which each option has its own implied volatility each day, and ad hoc-2 is the other ad hoc version, in which the implied volatility (on each day) for each option is estimated via the OLS procedure discussed previously.

The first panel clearly shows that judging models on the basis of hedging/replication errors could be somewhat different from judging them on the basis of valuation errors, as discussed previously; valuation errors could include either in-sample errors that show how well the model values fit market prices or out-of-sample/predictive error.[25] For example, ad hoc-2 yields substantially lower prediction errors than the Black-Scholes-Merton-2 version (Heston and Nandi forthcoming) but is the least competitive in terms of hedging errors. On the other hand, the magnitude of hedging errors of ad hoc-1, in which the in-sample valuation errors is essentially zero (as each option is priced exactly), is not very different from that of Black-Scholes-Merton-1. In fact, Black-Scholes-Merton-1, which has the highest in-sample valuation errors (as volatility is not

**TABLE 4 Mean Absolute Hedging Errors**

| | BSM-1 | BSM-2 | Ad Hoc-1 | Ad Hoc-2 |
|---|---|---|---|---|
| Whole Sample, All Options | | | | |
| One-day | $0.46 | $0.45 | $0.43 | $0.52 |
| Three-day | $0.66 | $0.65 | $0.62 | $0.78 |
| Five-day | $0.98 | $0.94 | $0.87 | $1.07 |
| Far-out-of-the-Money Puts under Forty Days to Maturity | | | | |
| One-day | $0.22 | $0.16 | $0.10 | $0.19 |
| Three-day | $0.23 | $0.19 | $0.20 | $0.26 |
| Five-day | $0.63 | $0.50 | $0.40 | $0.64 |
| Near-the-Money Calls under Forty Days to Maturity | | | | |
| One-day | $0.25 | $0.33 | $0.24 | $0.34 |
| Three-day | $0.49 | $0.52 | $0.44 | $0.60 |
| Five-day | $0.98 | $1.08 | $0.90 | $0.83 |
| Near-the-Money Puts Forty to Seventy Days to Maturity | | | | |
| One-day | $0.52 | $0.56 | $0.53 | $0.62 |
| Three-day | $0.74 | $0.76 | $0.77 | $0.91 |
| Five-day | $1.20 | $1.34 | $1.15 | $1.17 |

Source: Calculated by the Federal Reserve Bank of Atlanta using data from Standard and Poor's 500 index options market

BOX 3

# Data Description

The data set used for hedging is a subset of the tick-by-tick data on the Standard and Poor's 500 options that includes both the bid-ask quotes and the transaction prices; the raw data set is obtained directly from the exchange. The market for Standard and Poor's 500 index options is the second most active index options market in the United States, and in terms of open interest in options it is the largest. It is also easier to hedge Standard and Poor's 500 index options because there is a very active market for the Standard and Poor's 500 futures that are traded on the Chicago Mercantile Exchange.

Since many of the stocks in the Standard and Poor's 500 index pay dividends, a time series of dividends for the index is necessary. The daily cash dividends for the index collected from the Standard and Poor's 500 information bulletin for the years 1992–94 can be used.[1] The present value of the dividends (until the option expires) is computed and subtracted from the current index level. For the risk-free rate, the continuously compounded Treasury bill rates (from the average of the bid and ask discounts reported in the *Wall Street Journal*) are interpolated to match the maturity of the option.

The raw intraday data set is sampled every Wednesday (or the next trading day if Wednesday is a holiday) between 2:30 P.M. and 3:15 P.M. central standard time to create the data set.[2] In particular, given a particular Wednesday, an option must be traded on the following five trading days to be included in the sample. The study follows Dumas, Fleming, and Whaley (1998) in filtering the intraday data to create weekly data sets and use the midpoint of the bid-ask as the option price. As in Dumas, Fleming, and Whaley (1998), options with moneyness, $|K/F - 1|$ ($K$ is the strike price and $F$ is the forward price), less than or equal to 10 percent are included. In terms of maturity, options with time to maturity less than six days or greater than one hundred days are excluded.[3]

An option of a particular moneyness and maturity is represented only once in the sample on any particular day. In other words, although the same option may be quoted again in our time window (with the same or different index levels) on a given day, only the first record of that option is included in our sample for that day.

A transaction must satisfy the no-arbitrage relationship (Merton 1973) in that the call price must be greater than or equal to the spot price minus the present value of the remaining dividends and the discounted strike price. Similarly, the put price has to be greater than or equal to the present value of the remaining dividends plus the discounted strike price minus the spot price.

The entire data set consists of 7,404 records and observations spanning each trading day from January 5, 1994, to October 19, 1994.

1. Thanks to Jeff Fleming of Rice University for making the dividend series available.
2. Wednesdays are used as fewer holidays fall on Wednesdays.
3. See Dumas, Fleming, and Whaley (1998) for justification of the exclusionary criteria about moneyness and maturity.

20. If the number of options on a given day is too few, then there is a potential problem of overfitting in that more independent variables exist in the right-hand side but only a limited number of observations. However, such a problem can be partially mitigated by using a subset of the above regression (see Dumas, Fleming, and Whaley 1998).
21. One would want to test any options model in a very liquid options market so that prices are more reliable and do not reflect any liquidity premium.
22. The 1994 data were the latest full-year data available at the time of this writing.
23. The day $t$ is usually a Wednesday. If Wednesday is a holiday, then the next trading day is chosen.
24. The mean absolute hedging error is the mean of the absolute values of the hedging errors. The conclusions do not change if a slightly different criterion is used, like root mean squared hedging error.
25. Prediction or out-of-sample valuation errors measure how well a given model values options based on the model parameters that were estimated in a previous time period.

## GLOSSARY

**Call option:** Gives the owner of the option the right (but not the obligation) to buy the underlying asset at a fixed price (called the strike or exercise price). This right can be exercised at some fixed date in the future (European option) or at any time until the option matures (American option).

**Put option:** Gives the owner of the option the right (but not the obligation) to sell the underlying asset at a fixed price (called the strike or exercise price). This right can be exercised at some fixed date in the future (European option) or at any time until the option matures (American option).

**Long position:** In a security, implies that one has bought the security and currently owns it.

**Short position:** In a security, implies that one has sold a security that one does not own, but has only borrowed, with the hope of buying it back at a lower price in the future.

**Implied volatility:** The value of the volatility in the Black-Scholes-Merton formula that equates the model value of the option to its market price.

**In-sample errors:** Errors in fitting a model to data under a particular criterion function. For example, an options valuation model may have a few parameters or variables, the values of which are not observed directly. In such a case these parameters are estimated by minimizing a criterion function, such as the sum of squared differences between the model values and the market prices; this procedure is often called in-sample estimation. The differences between the model option values, evaluated at the estimates of the parameters, and the market option prices are called in-sample errors.

**Out-of-sample errors:** Measure the difference between the model option values and the market option prices on a sample of option prices that were observed at a later date than the sample on which the parameters of the model were estimated. In computing out-of-sample option values, the model parameters are fixed at the estimates obtained from the in-sample estimation.

implied but is computed from history of returns of the S&P 500 index), is quite competitive in terms of hedging across the entire sample of options.

Given the hedging results in the previous paragraph, which model would one choose among the four for constructing a hedge portfolio? The answer may very well depend on which option is to be hedged. The other panels of Table 4 show the mean absolute hedging errors of the four versions for three different classes of options: near-the-money call and put options and some relatively far-out-of-the-money put options. Most of these options are heavily traded in the Standard and Poor's 500 index options market.[26]

The table shows that the differences in hedging errors among most of the versions are more clearly manifested in far-out-of-the-money put options. The ad hoc-1 version, in which the delta of an option is computed from its exact implied volatility, clearly dominates in terms of hedging out-of-the-money puts, irrespective of the maturity. For near-the-money options, the differences between the various versions are not that significant, especially if the portfolio is rebalanced on the next day. In fact, the least complex of all the versions, Black-Scholes-Merton-1, is quite competitive in terms of hedging near-the-money options.

## Conclusion

Although the classic Black-Scholes-Merton paradigm of dynamic hedging is elegant from a theoretical perspective, it is often fraught with problems when it is implemented in the real world. Even if the Black-Scholes-Merton model were free of its known biases, the replicating/hedge portfolio of the model, which requires continuous trading, would rarely be able to match its target because trading can occur only at discrete intervals of time. Nevertheless, because of its simplicity and tractability, the model is widely used by options traders and investors. The basic delta-neutral hedge portfolio of the Black-Scholes-Merton model is also sometimes supplemented with other options to hedge a time-varying volatility (vega hedging). Although hedging a time-varying volatility is inconsistent with the Black-Scholes-Merton model, it can often prove useful in practice.

One would expect the presence of biases observed in the Black-Scholes-Merton model, such as the smile or smirk in implied volatilities, to result in further deterioration of the model's hedging performance. More advanced option pricing models (for example, random volatility models) that can account for some of the biases turn out to be useful mostly for deep out-of-the-money options but not

necessarily for near-the-money options. Ad hoc variations of the Black-Scholes-Merton model sometimes employed by options traders or investors to overcome the biases may also generate higher hedging errors than the very basic model despite the fact that ad hoc models often dominate the simple model in terms of matching observed option prices and predicting them. Although the simple Black-Scholes-Merton model can exhibit pricing biases, it is often competitive in terms of hedging because the pricing biases that it exhibits remain relatively stable through time.

Static hedging, an alternative to dynamic hedging, may seem promising because it is independent of any particular option pricing model. In particular, static hedging could prove useful for certain kinds of exotic options. However, static hedging requires hedging an option via other options so that the efficacy of static hedging depends on the liquidity of the options market, which often is not as liquid as the market on the underlying asset.

26. Far-out-of-the-money puts are those for which $K/F < 0.95$ where $K$ is the strike price and $F$ is the forward price for maturity $\tau$—that is, $F(t) = S(t)\exp[r(\tau)\tau]$, where $\tau$ is the time to maturity of the option. Near-the-money options are those for which $| K/F - 1 | \leq 0.01$.

## REFERENCES


BATES, DAVID. 1996. "Jumps and Stochastic Volatility: Exchange Rate Processes Implicit in Deutschemark Options." *Review of Financial Studies* 9 (Spring): 69–107.

BLACK, FISCHER. 1976. "Studies of Stock Price Volatility Changes." In *Proceedings of the 1976 Meetings of the American Statistical Association, Business and Economic Statistics Section*, 177–81. Alexandria, VA: American Statistical Association.

BLACK, FISCHER, AND MYRON S. SCHOLES. 1973. "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy* 81 (May/June): 637–54.

CARR, PETER. 1998. "Randomization and the American Put." *Review of Financial Studies* 11 (Fall): 327–43.

CHRISTIE, ANDREW A. 1982. "The Stochastic Behavior of Common Stock Variances: Value, Leverage, and Interest Rate Effects." *Journal of Financial Economics* 10 (December): 407–32.

COX, JOHN C., AND MARK RUBINSTEIN. 1985. *Options Markets*. Englewood Cliffs, N.J.: Prentice-Hall.

DERMAN, EMANUEL, AND IRAJ KANI. 1994. "Riding on the Smile." *Risk* 7 (February): 32–39.

DUMAS, BERNARD, JEFF FLEMING, AND ROBERT WHALEY. 1998. "Implied Volatility Functions: Empirical Tests." *Journal of Finance* 53 (December): 2059–2106.

DUPIRE, BRUNO. 1994. "Pricing with a Smile." *Risk* 7 (February): 18–20.

HARRISON, J. MICHAEL, AND STANLEY R. PLISKA. 1981. "Martingales and Stochastic Integrals in the Theory of Continuous Trading." *Stochastic Processes and Their Applications* 11:215–60.

HESTON, STEVEN L. 1993. "A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options." *Review of Financial Studies* 6, no. 2:327–43.

HESTON, STEVEN L., AND SAIKAT NANDI. Forthcoming. "A Closed-Form GARCH Option Valuation Model." *Review of Financial Studies*.

HUANG, JING-ZHI, MARTI G. SUBRAHMANYAM, AND G. GEORGE YU. 1996. "Pricing and Hedging American Options: A Recursive Integration Method." *Review of Financial Studies* 9 (Spring): 277–300.

HULL, JOHN C. 1997. *Options, Futures, and Other Derivatives*. 3d ed. Upper Saddle River, N.J.: Prentice-Hall.

JACKWERTH, JENS, AND MARK RUBINSTEIN. 1996. "Recovering Probability Distributions from Option Prices." *Journal of Finance* 51 (December): 1611–52.

JU, NENGJIU. 1998. "Pricing an American Option by Approximating Its Early Exercise Boundary as a Multipiece Exponential Function." *Review of Financial Studies* 11 (Fall): 627–46.

MERTON, ROBERT C. 1973. "The Theory of Rational Option Pricing." *Bell Journal of Economics* 4 (Spring): 141–83.

NANDI, SAIKAT. 1998. "How Important Is the Correlation between Returns and Volatility in a Stochastic Volatility Model? Empirical Evidence from Pricing and Hedging in the S&P 500 Index Options Market." *Journal of Banking and Finance* 22 (May): 589–610.

RUBINSTEIN, MARK. 1994. "Implied Binomial Trees." *Journal of Finance* 49 (July): 771–818.

VARIAN, HAL R. 1984. *Microeconomic Analysis*. New York: W.W. Norton and Company.

# Evidence on the Efficiency of Index Options Markets

**LUCY F. ACKERT AND YISONG S. TIAN**

*Ackert is a senior economist in the financial section of the Atlanta Fed's research department. Tian is an associate professor of finance at the Schulich School of Business, York University.* *The authors thank Jerry Dwyer, Mark Fisher, Saikat Nandi, and Steve Smith for helpful comments.*

SINCE THE CHICAGO BOARD OPTIONS EXCHANGE INTRODUCED THE FIRST INDEX OPTION CONTRACT IN 1983, INDEX OPTIONS MARKETS HAVE HAD A SIGNIFICANT ROLE IN FINANCIAL MARKETS. INDEX OPTIONS HAVE BEEN ONE OF THE MOST SUCCESSFUL OF THE MANY INNOVATIVE FINANCIAL INSTRUMENTS INTRODUCED OVER THE LAST FEW DECADES, AS THEIR HIGH TRADING volume indicates. Index options give market participants the ability to participate in anticipated market movements without having to buy or sell a large number of securities, and they permit portfolio managers to limit downside risk. Given their prominence and functions, the pricing efficiency of these markets is of great importance to academics, practitioners, and regulators.

Well-functioning financial markets are vital to a thriving economy because these markets facilitate price discovery, risk hedging, and allocating capital to its most productive uses. Inefficient pricing of index options indicates that their market (and, possibly, other financial markets) is not doing the best possible job at these important functions. To detect inefficient pricing (often called mispricing) requires computing a theoretically efficient price or price range and comparing it with prices of options traded in financial markets. But valuing an index option in theory is complicated and challenging.

One popular approach to deriving option pricing relationships is based on a principal called no-arbitrage. This approach is a very powerful tool in the valuation of financial assets because it does not make strong assumptions about traders' behavior or market price dynamics. The principle simply assumes that arbitrageurs enter the market and quickly eliminate mispricing if a riskless profit opportunity exists. An arbitrageur is an individual who takes advantage of a situation in which securities are mispriced relative to each other. The arbitrageur buys the underpriced asset and sells the overpriced asset, locking in a riskless profit. In doing so, the arbitrageur drives the price of the underpriced asset up and the price of the overpriced asset down, thus eliminating mispricing. However, in a well-functioning economy—where there are no free lunches—there is no portfolio of assets that has zero cost today and a certain, positive payoff in the future. Similarly, there is no port-

folio of assets that pays a positive amount with certainty today and requires no payment in the future. Arbitrage is critical for ensuring market efficiency because it forces asset prices to return to their implied, no-arbitrage values.

Many earlier studies report evidence of mispricing of index options, though arbitrage might have been limited. In some situations, market frictions restrict arbitrage so that investors simply cannot take advantage of available profit opportunities. For example, if arbitrageurs are subject to capital constraints and they cannot raise the capital necessary to form the riskless hedge, they will be unable to undertake trades that would move the market toward an efficient state (Shleifer and Vishny 1997). Similarly, the activity of arbitrageurs may be limited because the stock index underlying the option is often relatively difficult and costly to reproduce. To arbitrage based on a mispriced index option, investors may need to replicate the index by buying or selling a large basket or set of stocks that is perfectly correlated with the index. Doing so may be relatively difficult and costly, even for large investors (Ackert and Tian 1998b, 1999).[1]

The evidence of index option mispricing has been taken to indicate that options markets are inefficient and casts doubt on their contributions to price discovery, hedging, and efficient capital allocation. This article is a discussion of index option pricing aimed at analyzing earlier evidence of mispricing and presenting new evidence on index option pricing and its evolution. It first presents theoretical pricing relationships implied by no-arbitrage conditions. These conditions place bounds on possible efficient call and put option prices and imply relative pricing relationships between call and put option prices. A call (put) option is the option to buy (sell) an asset. Empirical tests of the conditions presented provide powerful insight into how options market efficiency has evolved over time. In contrast to many previous studies of options market efficiency, the arbitrage strategies examined here do not involve trading a stock index, and the relationships hold for any given value of the underlying asset. This approach avoids some of the difficulties that arise from impediments to arbitrage when, for example, an investor might have to short sell a large stock basket—that is, sell shares he or she does not own by borrowing them from another investor.

The article also reviews earlier studies of the pricing efficiency of index options markets and provides an empirical examination of the efficiency of the market for the popular Standard and Poor's (S&P) 500 index options. The results indicate some substantial deviations of market prices from theoretical pricing relationships. Importantly, S&P 500 index options are frequently mispriced, and the mispricing does not appear to have abated over time. The mispricing may not, however, indicate market inefficiency because there are limits to arbitrage.

**Index options give market participants the ability to participate in anticipated market movements without having to buy or sell a large number of securities, and they permit portfolio managers to limit downside risk.**

## Arbitrage Pricing Relationships

In evaluating the efficiency of option pricing, a theoretical optimal price derived from a model frequently provides the basis for comparison. Such theoretical models often assume specific dynamics for the underlying asset in order to derive more well-defined restrictions on the efficient price. In contrast, tests of pricing efficiency based solely on no-arbitrage arguments may be more informative if the relationships are independent of the models, though restrictions they place on price may not be very demanding.

Arbitrage pricing relationships are based on the simple assumption that investors prefer more to less. If these pricing relationships are violated by actual prices after adjustment for the bid-ask spread and transaction costs, arbitrage profits may be possible by buying the underpriced asset(s) and short-selling the overpriced asset(s). As discussed previously, rational pricing of options imposes explicit restrictions on the relative prices of call and put options. If these restrictions are violated, arbitrage opportunities exist. Some arbitrage pricing relationships jointly test options and stock market efficiency and allow examination of the information exchange between these markets whereas others test options market efficiency alone and allow examination of how pricing has evolved over time.[2] The relationships and

1. Note that traders can use the very liquid S&P 500 futures contract to replicate the index.
2. Billingsley and Chance (1985) and Ronn and Ronn (1989) note that some tests are joint tests of options and stock market efficiency while others consider only options market efficiency. See Ackert and Tian (1998a, 1999) for examples of both types of relationships.

empirical tests reported in this article are of the latter type. Because stock market transactions are not involved, examining these relationships may provide a superior test of pricing across index options. Another advantage of these types of relationships is that they are unaffected by the different closing times in stock and options markets.

The arbitrage pricing relationships presented below allow examining whether options market efficiency improved over the sample period. Options on the S&P 500 index are European, and the discussion below applies to European options only. A European option may be exercised only at maturity whereas an American option may be exercised prior to maturity. All relationships account for the bid-ask spread because bid-ask spreads result in significant transaction costs for participants in options markets (Phillips and Smith 1980; Baesel, Shows, and Thorp 1983). Define:

- $C^b$: bid price of European call option;
- $C^a$: ask price of European call option;
- $P^b$: bid price of a European put option;
- $P^a$: ask price of a European put option;
- $S$: price of underlying asset;
- $X$: strike price;
- $T$: maturity of the option;
- $r$: risk-free rate of interest or Treasury bill rate;[3]
- $t_i$: transaction costs (other than those arising from the bid-ask spread) of buying or selling calls, puts, or Treasury bills, $i = c$, $p$, or $r$.

Three sets of arbitrage pricing relationships are presented: the box spread, call and put spreads, and call and put convexity. The box spread is a combination of call and put spreads that matches two pairs of call and put options.[4] This strategy requires that an investor purchase and sell calls (bullish call spread) with strike prices $X_1$ and $X_2$, respectively, while simultaneously selling and purchasing puts (bearish put spread) with strike prices $X_1$ and $X_2$, respectively. The box spread is a riskless strategy because the future payoff is always positive: the difference between two strike prices, $X_2 - X_1$, where $X_1 < X_2$. The payoff is illustrated in Chart 1. If bid-ask spreads and transaction costs are taken into account, the box spread is expressed by the following two inequalities:

$$(C_1^a - C_2^b) - (P_1^b - P_2^a) + (X_1 - X_2)e^{-rT} + t_1 \geq 0 \quad (1a)$$

and

$$(C_2^a - C_1^b) - (P_2^b - P_1^a) + (X_2 - X_1)e^{-rT} + t_1 \geq 0, \quad (1b)$$

where $t_1 = 2t_c + 2t_p + t_r$. In the absence of arbitrage, inequalities (1a) and (1b) hold.

In contrast to the box spread, the call (put) spread combines two call (put) options with identical maturity. The call spread strategy requires purchase of call option 1 and sale of call option 2, where $X_1 < X_2$, as illustrated in Chart 2. The call spread is expressed as

$$(C_2^a - C_1^b) + (X_2 - X_1)e^{-rT} + t_{2a} \geq 0, \quad (2a)$$

where $t_{2a} = 2t_c + t_r$. Similarly, the put spread involves the sale of put option 1 and purchase of put option 2 and is expressed as

$$(P_1^a - P_2^b) + (X_2 - X_1)e^{-rT} + t_{2b} \geq 0, \quad (2b)$$

where $t_{2b} = 2t_p + t_r$.

Finally, call (put) convexity creates a riskless position by combining three call (put) options where $X_1 < X_2 < X_3$. The call (put) convexity strategy requires purchase of call (put) options 1 and 3 and sale of call (put) option 2. Call convexity is expressed as

$$wC_1^a + (1 - w)C_3^a - C_2^b + 2t_c \geq 0 \quad (3a)$$

and put convexity is

$$wP_1^a + (1 - w)P_3^a - P_2^b + 2t_p \geq 0, \quad (3b)$$

where $w = (X_3 - X_2)/(X_3 - X_1)$. So, for example, if $w = ½$, call convexity involves the purchase of one call option 1 and one call option 3 for every two call option 2s sold. The payoff from this strategy, commonly referred to as a butterfly spread, is illustrated in Chart 3.

If the box spread, call spread, put spread, call convexity, or put convexity is violated, arbitrage profits are possible by taking appropriate option positions. For example, if the call spread (2a) is violated, index call option 1 is overvalued relative to call option 2. The arbitrageur would sell call 1 and buy call 2, investing the balance in a Treasury bill earning the risk-free rate. In the case of exercise of both call options at maturity, the arbitrageur closes the index position and earns a risk-free profit at maturity (time $T$) of $(C_1^b - C_2^a) + (X_1 - X_2)e^{-rT} - t_{2a} \geq 0$, where $t_{2a} = 2t_c + t_r$ (see the box for the details of this arbitrage).

Although a violation of any of the inequalities above indicates the presence of an arbitrage opportunity, the box spread inequalities (1a) and (1b) place more demanding restrictions on the pricing of options. In the absence of transaction costs, the box spread requires an equality among four option prices. In contrast, even ignoring trans-

**CHART 1 Box Spread Payoff**

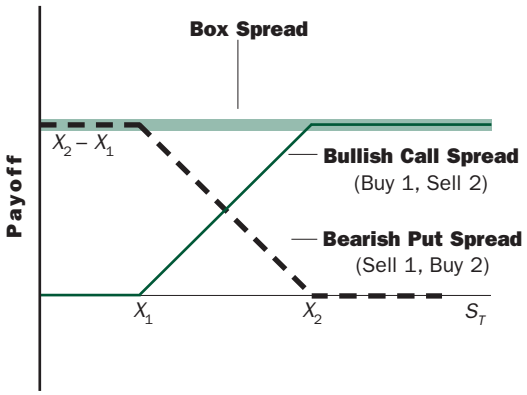



**CHART 2 Call Spread Payoff**

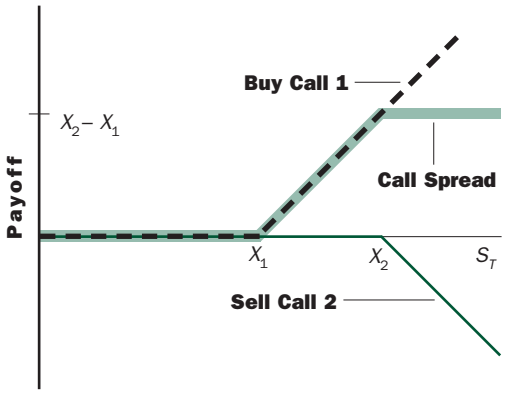
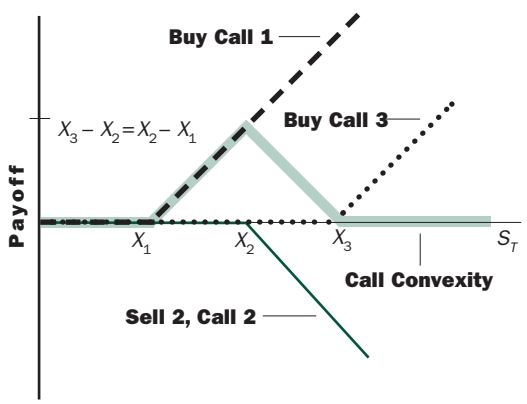



**CHART 3 Call Convexity ( $w = \frac{1}{2}$ )**




action costs, call and put spreads and convexity are minimum-maximum (inequality) restrictions, and a wide range of prices is consistent with the boundaries they place on option prices, as is apparent in Charts 1–3.

## Efficiency of Index Options Markets

Many empirical studies have tested pricing relations between put and call options, particularly for options on individual stocks. See, for example, Stoll (1969), Gould and Galai (1974), and Klemkosky and Resnick (1979). Some of these tests are based on theoretical option pricing models, such as the Black-Scholes (1973) or the Cox, Ross, and Rubinstein (1979) binomial option pricing models. Other tests are based simply on arbitrage arguments and are model-independent, including, for example, the box spread.

Although the empirical evidence generally supports some pricing relationships like put-call parity for individual stock options, significant mispricing has been reported in stock index options markets. For example, Evnine and Rudd (1985) use intraday data for a two-month period in 1984 and find frequent violations of boundary conditions and put-call parity for S&P 100 and Major Market index options, both of which are American options.[5] Evnine and Rudd further conclude that these options are significantly mispriced relative to theoretical values based on the binomial option pricing model. Chance (1987) also finds that put-call parity and the box spread are violated frequently for S&P 100 index options and that the violations are significant in size.[6] However, these results may not indicate market inefficiency for several reasons.

3. The assumption is that borrowing and lending rates are equal. Regarding the impact of this assumption on the results, see note 14.
4. The box spread is also a simple algebraic combination of the put-call parity relationship for each option. Put-call parity relates the put price, call price, exercise price, risk-free interest rate, and underlying asset price for options on the same asset with identical exercise price and expiration date. According to put-call parity, a pair of call and put options with identical maturity and strike price should be priced such that $C + Xe^{-rT} = P + S$, ignoring transaction costs and the bid-ask spread.
5. A boundary condition specifies a maximum or minimum price for an option. For example, an upper bound on the price of a call option is the value of the underlying asset because, no matter what happens, the option can never be worth more than the asset, that is, $C \leq S$. See note 4 above on put-call parity. For derivations of the various pricing relationships, see Merton (1973), Cox and Rubinstein (1985), Chance (1987), and Hull (1997).
6. In another study, Chance (1986) examines whether S&P 100 option prices are consistent with the Black-Scholes model and concludes that the model cannot be used to generate abnormal returns.

# Arbitrage Opportunity When the Call Spread Does Not Hold

Ignoring the bid-ask spread and transaction costs, the call spread is expressed as

$$(C_2 - C_1) + (X_2 - X_1)e^{-rT} \geq 0.$$

Without loss of generality, it is always possible to rearrange the pair of calls so that $X_1 < X_2$. Given this arrangement, the price of the first call should always be greater than that of the second call; that is, $C_1 > C_2$. If the call spread is violated, that is, if

$$(C_2 - C_1) + (X_2 - X_1)e^{-rT} < 0,$$

then risk-free profit opportunities are present. Arbitrage profits are possible by taking appropriate positions in the options market. In this case, index call option 1 is overvalued relative to call option 2. The arbitrageur would sell call 1, buy call 2, and invest the balance in a Treasury bill earning the risk-free rate. At maturity, a call option is exercised if the stock price exceeds its exercise price. The cash flows from the strategy are as shown in the table.

At the inception of this strategy there is no initial investment, and at maturity there are three possible cash flows, all of which are positive. When $S < X_1$, neither option is exercised upon maturity and the arbitrageur accrues the entire amount invested in the Treasury bill as profit; that is, $(C_1 - C_2)e^{rT} > 0$. When $X_1 < S \leq X_2$, option 1 is exercised but not option 2 and the investment in the Treasury bill more than offsets the loss on the first option so that

$$(C_1 - C_2)e^{rT} - (S - X_1) > (C_1 - C_2)e^{rT} - (X_2 - X_1) > 0.$$

The first inequality holds because $(X_1 - S) > (X_1 - X_2)$ when $X_1 < S \leq X_2$, and the second, because the call spread is violated. Finally, if both options are exercised at maturity, positive profit also accrues, again because the call spread is violated. Therefore, profit is made by the arbitrageur in all three possible outcomes. After commission fees and the bid-ask spread are recognized, the violation of (2a) (see Tables 4 and 5) must be large enough to compensate for transaction costs; that is,

$$(C_2^a - C_1^b) + (X_2 - X_1)e^{-rT} + t_{2a} < 0.$$

Such an opportunity cannot persist as arbitrageurs will take advantage of the mispricing until (2a) holds.

**Arbitrage When the Call Spread Is Violated**

| | Cash Flow | | | |
|---|---|---|---|---|
| | | At Options' Maturity | | |
| Strategy | Today | $S < X_1$ | $X_1 < S \leq X_2$ | $S > X_2$ |
| Sell call 1 | $+C_1$ | — | $-(S - X_1)$ | $-(S - X_1)$ |
| Sell call 2 | $-C_2$ | — | — | $S - X_2$ |
| Buy Treasury bill | $-(C_1 - C_2)$ | $(C_1 - C_2)e^{rT}$ | $(C_1 - C_2)e^{rT}$ | $(C_1 - C_2)e^{rT}$ |
| Total | 0 | $(C_1 - C_2)e^{rT}$ | $(C_1 - C_2)e^{rT} - (S - X_1)$ | $(C_1 - C_2)e^{rT} - (X_2 - X_1)$ |

These tests of market efficiency may be misleading because they use American options and the arbitrage conditions are for European options. Kamara and Miller (1995) point out that prior to their examination all tests of put-call parity used American options. In addition, tests of other arbitrage pricing relationships such as the box spread used data for American options (for example, Billingsley and Chance 1985). Because of the possibility of early exercise, these relationships may not be expected to hold for American options, and similar conditions for American options are frequently intractable. In their tests using S&P 500 index options that are European, Kamara and Miller find fewer and smaller violations.

Tests of put-call parity may also fail to indicate market inefficiency if arbitrage at low cost is not possible. Introducing index options helps to reduce arbitrage costs. In Canada, a stock basket, Toronto Index Participation Units (TIPS 35), has been traded since 1990. Ackert and Tian (1998a) examine the efficiency of Canadian index and options markets by comparing the number and size of violations in theoretical pricing relationships before and after the introduction of TIPs. They conclude that, although options market efficiency improved over their test period, the connection between options markets and stock markets did not. Ackert and Tian (1999) also examine the impact of a traded stock basket, Standard and Poor's Depositary Receipts (SPDRs), on the link between U.S. index options markets. They conclude that the introduction of a stock basket can enhance market efficiency because it removes one limit to arbitrage.

In summary, the results reported in earlier studies suggest that put-call parity is frequently violated in index options markets and that these options are often mispriced relative to prices predicted by theoretical models. To overcome problems in earlier studies, this study tests theoretical pricing relationships based on no-arbitrage conditions for European stock index options. It focuses on tests of options market efficiency independent of the stock market and includes the effects of transaction costs and bid-ask spreads. It also examines whether deviations from pricing relations declined over the 1986–96 sample period.

## Empirical Results

All three arbitrage pricing relationships presented earlier are investigated for S&P 500 index options on each trading day in the sample, as described subsequently. The number and size of violations are recorded and analyzed. This approach allows examining the evolution of the index options market and provides insight into whether market efficiency increased over the sample period. Arbitrage based on violations of the relationships considered does not require a position in the underlying asset. In addition, all of the pricing relationships are independent of an option pricing model so that no assumption concerning the process underlying the stock price is required. Thus, the empirical tests are true tests of market efficiency instead of joint tests of market efficiency and model specification.[7] Finally, the analysis recognizes the limits that transaction costs and bid-ask spreads place on arbitrage.

The empirical investigation analyzes the efficiency of the S&P 500 index options market using daily data for the S&P 500 index and index options from January 1, 1986, through December 31, 1996. Daily closing prices, trading volume, and open interest for S&P 500 index call and put options are from the Chicago Board Options Exchange.[8] The three-month Treasury bill rate (a proxy for the risk-free interest rate) is from the *Federal Reserve Bulletin*. Bid-ask spreads and commissions are included so that the analysis recognizes the effect of transaction costs on pricing efficiency. The approach is conservative in that it uses closing bid and ask prices, rather than closing prices, in testing the pricing relationships.[9] Following Harris, Sofianos, and Shapiro (1994) and Kamara and Miller (1995), this research constructs bid and ask prices, based on the usual spread in option prices, from closing prices. The option bid-ask spread is estimated by adding or subtracting ⅛32 (1/16) of a point if the price is less than (greater than or equal to) $3.[10] Following Kamara and Miller (1995), commission costs ($t_i$) are $30 for Treasury bills and $2 ($4) per option contract for 100 shares if the price is less than (greater than or equal to) $1.

On each trading day during the test period, the three pricing relationships discussed above are tested: the box spread (1a) and (1b), call and put spreads (2a) and (2b), and call and put convexity (3a) and (3b). For each maturity month, two pairs of put and call options are used to examine the box spread. The put and call within each pair are matched

7. So, for example, there is no test of whether prices are consistent with those predicted by a particular model such as the Black-Scholes option pricing model.
8. All relationships tested require synchronous option prices. Inferences are limited by the fact that closing prices may be nonsynchronous. However, Evnine and Rudd (1985) and Kamara and Miller (1995) find very similar results using intraday and closing price data for S&P 100 and S&P 500 index options, respectively.
9. See Ronn and Ronn (1989), who demonstrate that the use of bid-ask prices is conservative. They note that the market maker commits to transacting at least one contract at the bid-ask quotes, but the effective spread may be narrower. Traders are sometimes able to bargain to obtain better prices so that trades occur inside the quoted spread.
10. Some traders may have access to better price quotes. The assumption in this article concerning the constructed spread appears to be reasonable based on the results reported by others, though the results may be affected by the assumption to the extent that the spread is over- or underestimated.

with an identical strike price, but two different strike prices are used for the two pairs. In contrast, the call (put) spread combines two call (put) options with identical maturity and different strike prices. Finally, call (put) convexity combines three call (put) options with identical maturity and different strike prices.

The frequency and severity of violations are tabulated for the full sample period as well as for each year in the sample.[11] Examining violations in the pricing relationships for each sample year provides insight into how the efficiency of the options market has changed as the market has developed over time. Tables 1 through 9 report the percentage of violations as well as the mean violation in dollars. Significant dollar violations are tested for by testing the null hypothesis that the mean dollar violation is zero. All reported $t$ statistics use standard errors corrected for autocorrelation using a maximum likelihood procedure estimated by a Gauss-Marquardt algorithm (Judge and others 1985).[12]

To further investigate the persistence of violations in pricing relationships, the study examines whether arbitrage opportunities are evident the day following observed violations. Doing so provides an ex ante test, which, as Galai (1977) argues, a true test of market efficiency must be. Ex ante tests are executed from the trader's point of view and reflect the trader's ability to actually form the required, profitable portfolio. In an ex ante approach, current prices reveal arbitrage opportunities but execution is at prices that are yet to be revealed. Conducting ex ante tests involves identifying each day on which a particular violation occurs and tracking whether the violation persisted on the following trading day. Existence on the following day implies that traders did not fully eliminate arbitrage opportunities.

Table 1 reports the frequency and severity of violations and ex ante violations of the box spread, inequalities (1a) and (1b). For the two inequalities, the percentage and dollar amount of violations are similar (21.02 percent and $1.07 versus 23.78 percent and $1.08). For each relationship, the percentage of violations is substantial and the mean dollar violation is significantly different from zero.[13] The ex ante tests indicate that significant abnormal profit opportunities existed even on the day following the violation of a pricing relationship. For example, 28,292 violations of (1a) occurred, and of these violations 2,785 or 9.84 percent persisted on the following day with a significant mean violation of $1.02.[14]

Tables 2 and 3 report the percentage and dollar size of violations and ex ante violations of (1a) and (1b), respectively, for each year in the 1986–96 sample period. All mean dollar violations are significantly different from zero at the 1 percent significance level. Although some variation is observed in the extent to which the pricing relationships are violated across years, the results provide no evidence that options market efficiency improved over the sample

**TABLE 1 Violations and Ex Ante Violations of the Box Spread (1a) and (1b)**

| | Box Spread (1a) | | Box Spread (1b) | |
|---|---|---|---|---|
| | Total Violations | Ex Ante Violations | Total Violations | Ex Ante Violations |
| **Frequency of Violations** | | | | |
| Number of Observations | 134,606 | 28,292 | 134,606 | 32,014 |
| Number of Violations | 28,292 | 2,785 | 32,014 | 3,210 |
| Percentage of Violations | 21.02 | 9.84 | 23.78 | 10.03 |
| **Violations, in Dollars** | | | | |
| Mean | 1.07 | 1.02 | 1.08 | 1.11 |
| Standard Deviation | 1.05 | 1.00 | 1.07 | 1.09 |
| $t$ statistic for nonzero mean | 170.00*** | 54.19*** | 180.36*** | 57.74*** |

Note: This table reports the frequency and dollar size of violations of the box spread (1a) and (1b) using daily data for the S&P 500 index and index options from January 1, 1986, through December 31, 1996. An ex ante violation occurs when a particular violation persists into the following trading day. Asterisks *,**, or *** denote significance at the 10 percent, 5 percent, and 1 percent levels, respectively, in a two-tailed test.

**TABLE 2 Violations and Ex Ante Violations of the Box Spread (1a) by Year**

| | Total Violations | | Ex Ante Violations | |
|---|---|---|---|---|
| Sample Year | Percentage of Violations | Mean Dollar Violation | Percentage of Violations | Mean Dollar Violation |
| 1986 | 18.72 | 0.83 | 4.35 | 0.76 |
| 1987 | 22.30 | 1.24 | 8.03 | 1.13 |
| 1988 | 20.34 | 0.86 | 5.83 | 0.85 |
| 1989 | 13.94 | 0.91 | 7.54 | 0.81 |
| 1990 | 24.06 | 1.06 | 12.18 | 1.01 |
| 1991 | 21.62 | 0.99 | 9.11 | 0.90 |
| 1992 | 17.03 | 0.78 | 9.21 | 0.92 |
| 1993 | 18.29 | 0.86 | 9.42 | 0.76 |
| 1994 | 17.69 | 0.92 | 8.99 | 0.87 |
| 1995 | 20.76 | 1.02 | 10.75 | 1.05 |
| 1996 | 25.96 | 1.35 | 11.14 | 1.23 |
| Overall | 21.02 | 1.07 | 9.84 | 1.02 |

Note: This table reports the percentage and dollar size of violations of the box spread (1a) using daily data for the S&P 500 index and index options for each year in the January 1, 1986, through December 31, 1996, sample period. An ex ante violation occurs when a particular violation persists into the following trading day. All mean dollar violations are significantly different from zero at the 1 percent significance level.

period. The frequency of violations remains high at approximately 20 percent of observations, even after taking into account trading costs, including the bid-ask spread and commission fees.

Next, violations of call and put spreads (2a) and (2b) and call and put convexity (3a) and (3b) are examined. As reported in Tables 4 and 7, significant mean dollar violations and ex ante dollar violations were observed for all four relationships. However, for all four the frequency of violations is quite low. The maximum percentage of violations (ex ante violations) across the four inequalities for the full sample is only 3.08 percent (8.04 percent). When the percentage and dollar violations by year reported in Tables 5 and 6 (8 and 9) for call and put spreads (convexity) are considered, there is no apparent trend. Although market efficiency does not appear to have improved over the sample period, the results suggest that options market valuations were generally consistent with these theoretical predictions.

A numerical example for the call spread provides perspective on the size of the violations reported in this article. On January 4, 1996, call options expiring on March 16, 1996, with strike prices 610 ($X_1$) and 615 ($X_2$) were priced at \$23.25 ($C_1$) and \$15.50 ($C_2$). The maturity date translates into a time to maturity of 0.1973 years ($T$), and the continuously compounded Treasury bill rate is 5.29 percent ($r$). Using inequality (2a) and ignoring transaction costs results in $15.50 – 23.25 + (615 – 610)e^{(-0.0529 \times 0.1973)} = –2.8019$ so that the size of the violation is \$2.80. Transaction costs are the sum of commission fees and the bid-ask spread and are (4 + 4 + 30)/100 + 1/8 = 0.505, which gives a net violation of \$2.30 (–2.8019 + 0.505).

11. In some cases, a few extreme outliers were detected. After checking and rechecking the original data sources, these outliers remained. However, removing these outliers does not change statistical inferences.
12. Autocorrelation in the dollar violations might be expected because the time to maturity for sample options may overlap. Diagnostic tests confirm the presence of significant positive autocorrelation. However, inferences are unchanged if ordinary least squares standard errors are used.
13. Note that inequality (1a) involves lending whereas inequality (1b) requires borrowing. Because similar frequency and magnitude of violations are observed across the two inequalities, the results suggest that the assumption of equal borrowing and lending rates does not explain the extent of profit opportunities.
14. Abnormal profit opportunities are not expected to persist and, thus, the mean ex post violation is expected to be zero. However, no directional relationship in the percentage of violations over the two-day time period is posited.

## TABLE 3 Violations and Ex Ante Violations of the Box Spread (1b) by Year

| | Total Violations | | Ex Ante Violations | |
|---|---|---|---|---|
| Sample Year | Percentage of Violations | Mean Dollar Violation | Percentage of Violations | Mean Dollar Violation |
| 1986 | 21.32 | 0.83 | 7.07 | 0.60 |
| 1987 | 26.23 | 1.27 | 11.16 | 1.26 |
| 1988 | 21.56 | 0.95 | 5.65 | 0.67 |
| 1989 | 15.90 | 0.90 | 6.36 | 0.86 |
| 1990 | 25.11 | 1.03 | 8.99 | 1.12 |
| 1991 | 24.16 | 0.99 | 9.28 | 1.03 |
| 1992 | 20.25 | 0.88 | 8.45 | 0.84 |
| 1993 | 19.19 | 0.81 | 6.77 | 0.67 |
| 1994 | 19.73 | 0.91 | 8.89 | 1.04 |
| 1995 | 23.87 | 0.95 | 10.61 | 0.83 |
| 1996 | 30.54 | 1.41 | 13.17 | 1.42 |
| Overall | 23.78 | 1.08 | 10.03 | 1.11 |

Note: This table reports the percentage and dollar size of violations of the box spread (1b) using daily data for the S&P 500 index and index options for each year in the January 1, 1986, through December 31, 1996, sample period. An ex ante violation occurs when a particular violation persists into the following trading day. All mean dollar violations are significantly different from zero at the 1 percent significance level.

## TABLE 4 Violations and Ex Ante Violations of the Call Spread (2a) and Put Spread (2b)

| | Call Spread (2a) | | Put Spread (2b) | |
|---|---|---|---|---|
| | Total Violations | Ex Ante Violations | Total Violations | Ex Ante Violations |
| | Frequency of Violations | | | |
| Number of Observations | 283,345 | 5,806 | 537,701 | 2,159 |
| Number of Violations | 5,806 | 467 | 2,159 | 145 |
| Percentage of Violations | 2.05 | 8.04 | 0.40 | 6.72 |
| | Violations, in Dollars | | | |
| Mean | 1.05 | 1.09 | 1.30 | 1.08 |
| Standard Deviation | 1.04 | 1.06 | 1.22 | 1.06 |
| *t* statistic for nonzero mean | 77.24*** | 22.13*** | 49.47*** | 12.27*** |

Note: This table reports the frequency and dollar size of violations of the call spread (2a) and put spread (2b) using daily data for the S&P 500 index and index options from January 1, 1986, through December 31, 1996. An ex ante violation occurs when a particular violation persists into the following trading day. Asterisks *,**, or *** denote significance at the 10 percent, 5 percent, and 1 percent levels, respectively, in a two-tailed test.

## TABLE 5 Violations and Ex Ante Violations of the Call Spread (2a) by Year

| | Total Violations | | Ex Ante Violations | |
|---|---|---|---|---|
| Sample Year | Percentage of Violations | Mean Dollar Violation | Percentage of Violations | Mean Dollar Violation |
| 1986 | 1.08 | 0.81 | 7.69 | 1.72 |
| 1987 | 2.10 | 0.92 | 8.70 | 0.70 |
| 1988 | 1.14 | 0.85 | 2.69 | 1.95 |
| 1989 | 1.71 | 0.91 | 6.54 | 0.86 |
| 1990 | 0.80 | 0.85 | 7.07 | 1.11 |
| 1991 | 3.14 | 1.17 | 8.36 | 0.90 |
| 1992 | 1.64 | 0.98 | 7.51 | 1.21 |
| 1993 | 1.48 | 0.73 | 6.65 | 0.52 |
| 1994 | 0.70 | 0.88 | 2.46 | 0.91 |
| 1995 | 3.92 | 0.99 | 11.16 | 1.14 |
| 1996 | 2.21 | 1.36 | 6.18 | 1.47 |
| Overall | 2.05 | 1.05 | 8.04 | 1.09 |

Note: This table reports the percentage and dollar size of violations of the call spread (2a) using daily data for the S&P 500 index and index options for each year in the January 1, 1986, through December 31, 1996, sample period. An ex ante violation occurs when a particular violation persists into the following trading day. All mean dollar violations are significantly different from zero at the 1 percent significance level.

## TABLE 6 Violations and Ex Ante Violations of the Put Spread (2b) by Year

| | Total Violations | | Ex Ante Violations | |
|---|---|---|---|---|
| Sample Year | Percentage of Violations | Mean Dollar Violation | Percentage of Violations | Mean Dollar Violation |
| 1986 | 0.31 | 0.97 | 0 | 0 |
| 1987 | 1.83 | 1.75 | 4.75 | 1.66 |
| 1988 | 0.48 | 0.81 | 3.91 | 0.37 |
| 1989 | 0.19 | 0.90 | 3.70 | 0.25 |
| 1990 | 0.87 | 1.13 | 8.00 | 0.96 |
| 1991 | 0.25 | 0.97 | 4.42 | 0.43 |
| 1992 | 0.24 | 0.79 | 7.29 | 0.74 |
| 1993 | 0.17 | 0.96 | 2.78 | 0.19 |
| 1994 | 0.45 | 1.05 | 9.51 | 0.88 |
| 1995 | 0.12 | 1.00 | 9.80 | 1.06 |
| 1996 | 0.29 | 1.65 | 8.06 | 1.34 |
| Overall | 0.40 | 1.30 | 6.72 | 1.08 |

Note: This table reports the percentage and dollar size of violations of the put spread (2b) using daily data for the S&P 500 index and index options for each year in the January 1, 1986, through December 31, 1996, sample period. An ex ante violation occurs when a particular violation persists into the following trading day. All mean dollar violations are significantly different from zero at the 1 percent significance level.

## TABLE 7 Violations and Ex Ante Violations of Call Convexity (3a) and Put Convexity (3b)

| | Call Convexity (3a) | | Put Convexity (3b) | |
|---|---|---|---|---|
| | Total Violations | Ex Ante Violations | Total Violations | Ex Ante Violations |
| | Frequency of Violations | | | |
| Number of Observations | 882,954 | 27,206 | 2,244,467 | 20,439 |
| Number of Violations | 27,206 | 1,659 | 20,439 | 844 |
| Percentage of Violations | 3.08 | 6.10 | 0.91 | 4.13 |
| | Violations, in Dollars | | | |
| Mean | 0.91 | 1.13 | 0.95 | 1.21 |
| Standard Deviation | 0.94 | 1.07 | 1.04 | 1.14 |
| *t* statistic for nonzero mean | 159.98*** | 43.01*** | 131.12*** | 30.81*** |

Note: This table reports the frequency and dollar size of violations of call convexity (3a) and put convexity (3b) using daily data for the S&P 500 index and index options from January 1, 1986, through December 31, 1996. An ex ante violation occurs when a particular violation persists into the following trading day. Asterisks *,**, or *** denote significance at the 10 percent, 5 percent, and 1 percent levels, respectively, in a two-tailed test.

## TABLE 8 Violations and Ex Ante Violations of Call Convexity (3a) by Year

| | Total Violations | | Ex Ante Violations | |
|---|---|---|---|---|
| Sample Year | Percentage of Violations | Mean Dollar Violation | Percentage of Violations | Mean Dollar Violation |
| 1986 | 2.11 | 0.72 | 0.61 | 0.02 |
| 1987 | 3.83 | 0.99 | 9.82 | 1.29 |
| 1988 | 1.37 | 0.79 | 1.12 | 0.20 |
| 1989 | 2.23 | 0.76 | 2.29 | 0.80 |
| 1990 | 1.76 | 0.64 | 4.07 | 0.72 |
| 1991 | 4.32 | 1.00 | 8.41 | 1.09 |
| 1992 | 1.82 | 0.75 | 3.17 | 1.02 |
| 1993 | 1.66 | 0.54 | 4.00 | 0.36 |
| 1994 | 0.85 | 0.55 | 4.09 | 0.82 |
| 1995 | 4.98 | 0.85 | 6.65 | 1.05 |
| 1996 | 3.45 | 1.07 | 5.12 | 1.33 |
| Overall | 3.08 | 0.91 | 6.10 | 1.13 |

Note: This table reports the percentage and dollar size of violations of call convexity (3a) using daily data for the S&P 500 index and index options for each year in the January 1, 1986, through December 31, 1996, sample period. An ex ante violation occurs when a particular violation persists into the following trading day. All mean dollar violations are significantly different from zero at the 1 percent significance level.

**TABLE 9 Violations and Ex Ante Violations of Put Convexity (3b) by Year**

| | Total Violations | | Ex Ante Violations | |
|---|---|---|---|---|
| Sample Year | Percentage of Violations | Mean Dollar Violation | Percentage of Violations | Mean Dollar Violation |
| 1986 | 1.06 | 0.58 | 0 | 0 |
| 1987 | 4.21 | 1.53 | 6.31 | 1.76 |
| 1988 | 0.77 | 0.71 | 2.06 | 0.49 |
| 1989 | 0.40 | 0.64 | 1.03 | 0.13 |
| 1990 | 1.82 | 0.82 | 5.59 | 0.93 |
| 1991 | 0.56 | 0.76 | 2.00 | 0.66 |
| 1992 | 0.45 | 0.70 | 2.91 | 0.70 |
| 1993 | 0.20 | 0.56 | 0.36 | 0.33 |
| 1994 | 0.76 | 0.78 | 3.94 | 1.13 |
| 1995 | 0.29 | 0.52 | 2.05 | 0.36 |
| 1996 | 1.01 | 0.89 | 3.51 | 1.06 |
| Overall | 0.91 | 0.95 | 4.13 | 1.21 |

Note: This table reports the percentage and dollar size of violations of put convexity (3b) using daily data for the S&P 500 index and index options for each year in the January 1, 1986, through December 31, 1996, sample period. An ex ante violation occurs when a particular violation persists into the following trading day. All mean dollar violations are significantly different from zero at the 1 percent significance level.

Taken together, significant violations of arbitrage pricing relationships are observed, even using ex ante tests, particularly for the box spread relationship. The differing results across the relationships tested are not surprising because the box spread is a more demanding test of market efficiency as compared with call and put spreads or convexity. The overall finding is that S&P 500 index options are frequently mispriced to a significant extent and that options market efficiency has not changed markedly over time.

## Conclusion

This article examines the efficiency of the S&P 500 index options market using theoretical pricing relationships derived from stock index option no-arbitrage principles. It reports frequent and substantial violations of the box spread relationship in particular, even though the analysis reflects transaction costs. The results do not provide support for the argument that options market efficiency improved over time. However, at the same time, there were few violations of call and put spreads and convexity, which are less demanding tests of pricing efficiency than the box spread.

Market frictions appear to have a significant effect on arbitrageurs' abilities to take advantage of violations of no-arbitrage pricing relationships. Although the analysis reflects the market frictions imposed by the bid-ask spread and commission costs, other frictions may be significant. One such friction may be insufficient liquidity, which increases option traders' risk and may prevent them from eliminating arbitrage opportunities. In a liquid market a transaction can be quickly completed with little impact on prices.

## REFERENCES


ACKERT, LUCY F., AND YISONG S. TIAN. 1998a. "The Introduction of Toronto Index Participation Units and Arbitrage Opportunities in the Toronto 35 Index Option Market." *Journal of Derivatives* 5, no. 4:44–53.

———. 1998b. "Investor Sentiment and Mispricing in Traded Stock Portfolios." Unpublished paper, Federal Reserve Bank of Atlanta.

———. 1999. "Efficiency in Index Options Markets and Trading in Stock Baskets." Federal Reserve Bank of Atlanta Working Paper 99-5, June.

BAESEL, JEROME B., GEORGE SHOWS, AND EDWARD THORP. 1983. "The Cost of Liquidity Services in Listed Options." *Journal of Finance* 38 (June): 989–95.

BILLINGSLEY, RANDALL S., AND DON M. CHANCE. 1985. "Options Market Efficiency and the Box Spread Strategy." *Financial Review* 20 (November): 287–301.

BLACK, FISCHER, AND MYRON S. SCHOLES. 1973. "The Pricing of Options and Corporate Liabilities." *Journal of Political Economy* 81 (May/June): 637–54.

CHANCE, DON M. 1986. "Empirical Tests of the Pricing of Index Call Options." In *Advances in Futures and Options Research*. Vol. 1. Greenwich, Conn.: JAI Press.

———. 1987. "Parity Tests of Index Options." In *Advances in Futures and Options Research*. Vol. 2. Greenwich, Conn.: JAI Press.

COX, JOHN C., STEPHEN A. ROSS, AND MARK RUBINSTEIN. 1979. "Option Pricing: A Simplified Approach." *Journal of Financial Economics* 7 (September): 229–63.

COX, JOHN C., AND MARK RUBINSTEIN. 1985. *Options Markets*. Englewood Cliffs, N.J.: Prentice-Hall.

EVNINE, JEREMY, AND ANDREW RUDD. 1985. "Index Options: The Early Evidence." *Journal of Finance* 40 (July): 743–56.

GALAI, DAN. 1977. "Tests of Market Efficiency of the Chicago Board Options Exchange." *Journal of Business* 50 (April): 167–97.

GOULD, JOHN P., AND DAN GALAI. 1974. "Transactions Costs and the Relationship between Put and Call Prices." *Journal of Financial Economics* 1 (July): 105–29.

HARRIS, LAWRENCE, GEORGE SOFIANOS, AND JAMES E. SHAPIRO. 1994. "Program Trading and Intraday Volatility." *Review of Financial Studies* 7 (Winter): 653–85.

HULL, JOHN C. 1997. *Options, Futures, and Other Derivative Securities*. 3d ed. Upper Saddle River, N.J.: Prentice-Hall.

JUDGE, GEORGE G., W.E. GRIFFITHS, R. CARTER HILL, HELMUT LUKTEPOHL, AND TSOUNG-CHAO LEE. 1985. *The Theory and Practice of Econometrics*. 2d ed. New York: John Wiley and Sons.

KAMARA, AVRAHAM, AND THOMAS W. MILLER JR. 1995. "Daily and Intradaily Tests of European Put-Call Parity." *Journal of Financial and Quantitative Analysis* 30 (December): 519–39.

KLEMKOSKY, ROBERT C., AND BRUCE G. RESNICK. 1979. "Put-Call Parity and Market Efficiency." *Journal of Finance* 34 (December): 1141–55.

MERTON, ROBERT C. 1973. "The Theory of Rational Option Pricing." *Bell Journal of Economics* 4 (Spring): 141–83.

PHILLIPS, SUSAN M., AND CLIFFORD W. SMITH JR. 1980. "Trading Costs for Listed Options: The Implications for Market Efficiency." *Journal of Financial Economics* 8 (June): 179–201.

RONN, AIMEE GERBARG, AND EHUD I. RONN. 1989. "The Box Spread Arbitrage Conditions: Theory, Tests, and Investment Strategies." *Review of Financial Studies* 2, no. 1:91–108.

SHLEIFER, ANDREI, AND ROBERT W. VISHNY. 1997. "The Limits of Arbitrage." *Journal of Finance* 52 (March): 35–55.

STOLL, HANS R. 1969. "The Relationship between Put and Call Option Prices." *Journal of Finance* 24 (December): 801–24.