Methods for Evaluating Value-at-Risk Estimates

Jose A. Lopez

Economist, Federal Reserve Bank of San Francisco. The author would like to thank Beverly Hirtle, J.S. Butler, Bruce Hansen, Mark Levonian, Jim O'Brien, Joe Mattey, Jim Moser, Philip Strahan, and Egon Zakrajsek, as well as participants at the 1997 Federal Reserve System Conference on Financial Structure and Regulation for their comments.

Since 1998, U.S. commercial banks with significant trading activities have been required to hold capital against their defined market risk exposure. Under the "internal models" approach embodied in the current regulatory guidelines, the capital charges are a function of banks' own value-at-risk (VaR) estimates. VaR estimates are simply forecasts of the maximum portfolio loss that could occur over a given holding period with a specified confidence level. Clearly, the accuracy of these VaR estimates is of concern to both banks and their regulators.

To date, two hypothesis-testing methods for evaluating VaR estimates have been proposed, namely, the binomial and the interval forecast methods. For these tests, the null hypothesis is that the VaR estimates in question exhibit a specified property that is characteristic of accurate VaR estimates. As shown in a simulation exercise, these tests generally have low power and are thus prone to misclassifying inaccurate VaR estimates as "acceptably accurate."

An alternative evaluation method, based on regulatory loss functions, is proposed. Magnitude loss functions that assign quadratic numerical scores when observed portfolio losses exceed VaR estimates are shown to be particularly useful. Simulation results indicate that the loss function evaluation method is capable of distinguishing between VaR estimates generated by accurate and alternative VaR models. The additional information provided by this method as well as its flexibility with respect to the specification of the loss function make a reasonable case for its use in the regulatory evaluation of VaR estimates.

In August 1996, U.S. bank regulatory agencies adopted the market risk amendment (MRA) to the Basle Capital Accord. The MRA, which became effective in 1998, requires that commercial banks with significant trading activities set aside capital to cover the market risk exposure in their trading accounts.[1] The market risk capital requirements are based on the "value-at-risk" (VaR) estimates generated by the banks' own risk management models. In general, VaR models attempt to forecast the time-varying distributions of portfolio returns, and VaR estimates are simply specified lower quantiles of these forecasted distributions. In other words, VaR estimates are forecasts of the maximum portfolio loss that could occur over a given holding period with a specified confidence level.

Given the importance of VaR estimates to banks and now to their regulators, evaluating the accuracy of the models underlying them is a necessary exercise. According to Hendricks and Hirtle (1997):

> *The actual benefits to be derived from the VaR estimates depend crucially on the quality and accuracy of the models on which the estimates are based. To the extent that these models are inaccurate and misstate banks' true risk exposures, then the quality of the information derived from any public disclosure will be degraded. More important, inaccurate VaR models or models that do not produce consistent estimates over time will undercut the main benefit of a models-based capital requirement: the closer tie between capital requirements and true risk exposures. Thus, assessment of the accuracy of these models is a key concern and challenge for supervisors. (pp. 8–9)*

To date, two hypothesis-testing methods for evaluating VaR estimates have been proposed: the binomial method, the quantitative standard currently embodied in the MRA, and the interval forecast method proposed by Christoffersen (1998).[2] For these tests, the null hypothesis is that the VaR

1. Further details on the MRA are provided in Section I below. For complete details on the MRA, see the Federal Register (1996).

2. Note that other methods for evaluating VaR models have been proposed, but they focus on other aspects of the models' forecasted distributions. For example, Crnkovic and Drachman (1996) focus on the entire forecasted distribution, and Lopez (1999) focuses on probability forecasts generated from the forecasted distributions.

estimates in question exhibit a specified property that is characteristic of accurate VaR estimates. If the null hypothesis is rejected, the VaR estimates do not exhibit the specified property, and the underlying VaR model can be said to be "inaccurate." If the null hypothesis is not rejected, then the model can be said to be "acceptably accurate."

For these evaluation methods, as for any hypothesis test, a key issue is their statistical power, i.e., their ability to reject the null hypothesis when it is actually incorrect. If the hypothesis tests exhibit low power, then the probability of misclassifying an inaccurate VaR model as "acceptably accurate" will be high. A simulation exercise based on several data-generating processes finds the power of these tests to be quite low, thus limiting their usefulness for evaluating VaR estimates.

As an alternative to the hypothesis-testing framework, I propose an evaluation method that uses standard forecast evaluation techniques; that is, the accuracy of VaR estimates is gauged by how well they minimize a loss function that represents the evaluator's concerns. I consider three loss functions that represent specific regulatory concerns: (1) the binomial loss function that assigns a numerical score of 1 when a VaR estimate is exceeded by its corresponding portfolio loss, (2) the zone loss function based on the adjustments to the multiplication factor used in the MRA, and (3) the magnitude loss function that assigns a quadratic numerical score when a VaR estimate is exceeded by its corresponding portfolio loss.

Although statistical power is less relevant for this evaluation method, the related issues of comparative accuracy and model misclassification are examined within the context of a simulation exercise. The simulation results indicate that the degree of model misclassification generally mirrors that of the other methods. However, in certain cases, it provides additional useful information on the accuracy of VaR estimates. Of the three loss functions examined, the magnitude loss function seems to be more capable of distinguishing between accurate and alternative VaR estimates because it incorporates additional information—the magnitude of the trading losses—into the evaluation. The ability to use such additional information, as well as the flexibility with respect to the specification of the loss function, make a reasonable case for using the loss function method in the regulatory evaluation of VaR estimates.

Section I describes the current regulatory environment and the three evaluation methods. Section II presents the simulation results that indicate the usefulness of the proposed evaluation method, particularly using the magnitude loss function. Section III presents a detailed example of how this method can provide additional information useful in the regulatory evaluation of VaR estimates, and Section IV concludes.

I. ALTERNATIVE EVALUATION METHODS

VaR models are characterized by their forecasted distributions of k-period-ahead portfolio returns. To fix notation, let Y_t represent portfolio value at time t in dollar terms, and let $y_t = \ln(Y_t)$. The k-period-ahead portfolio return is $\varepsilon_{t+k} = y_{t+k} - y_t$. Conditional on the information available at time t, ε_{t+k} is a random variable with distribution f_{t+k}; that is, $\varepsilon_{t+k} \mid \Omega_t \sim f_{t+k}$. Thus, VaR model m is characterized by f_{mt+k}, its forecast of f_{t+k}.

VaR estimates are the most common type of forecast generated from VaR models. A VaR estimate is simply a specified quantile of the forecasted return distribution over a given holding period. The VaR estimate at time t derived from model m for a k-period-ahead return, denoted $\text{VaR}_{mt}(k,\alpha)$, is the critical value that corresponds to the lower α percent tail of f_{mt+k}. Thus, $\text{VaR}_{mt}(k,\alpha) = F^{-1}_{mt+k}(\alpha/100)$, where F^{-1}_{mt+k} is the inverse of the cumulative distribution function corresponding to f_{mt+k} or, equivalently, $\text{VaR}_{mt}(k,\alpha)$ is the solution to

$$\int_{-\infty}^{\text{VaR}_{mt}(k,\alpha)} f_{mt+k}(x)\,dx = \frac{\alpha}{100}.$$

Note that a VaR estimate is typically expressed in dollar terms as the loss between the current portfolio value and the portfolio value corresponding to it; that is, $\text{VaR}_{mt}(k,\alpha)$ is expressed in dollar terms as $\text{VaR\$}_{mt}(k,\alpha) = Y_t(1 - e^{\text{VaR}_{mt}(k,\alpha)})$.

Current Regulatory Framework

The U.S. capital rules for the market risk exposure of large commercial banks, effective as of 1998, are explicitly based on VaR estimates. The rules cover a bank's total trading activity, which is all assets in a bank's trading account (i.e., assets carried at their current market value) as well as all foreign exchange and commodity positions wherever located in the bank. Any bank or bank holding company whose total trading activity accounts for more than 10 percent of its total assets or is more than $1 billion must hold regulatory capital against its market risk exposure. The capital charge is calculated using the so-called "internal models" approach.

Under this approach, capital charges are based on VaR estimates generated by banks' internal risk management models using the standardizing parameters of a ten-day holding period ($k = 10$) and 99 percent coverage ($\alpha = 1$). In other words, a bank's market risk capital charge is based on its own estimate of the potential loss that would not be exceeded with 1 percent probability over the subsequent two-week period. The market risk capital that bank m must hold for time $t + 1$, denoted MRC_{mt+1}, is set as the larger of the dollar value of $\text{VaR}_{mt}(10,1)$ or a multiple of the average of the previous 60 $\text{VaR}_{mt}(10,1)$ estimates in dollar terms;

that is,

$$MRC_{mt+1} = \max\left[\text{VaR\$}_{mt}(10,1);\ S_{mt} * \frac{1}{60}\sum_{i=0}^{59}\text{VaR\$}_{mt-i}(10,1)\right] + SR_{mt},$$

where S_{mt} and SR_{mt} are a multiplication factor and an additional capital charge for the portfolio's idiosyncratic credit risk, respectively. Note that, under the current framework, $S_{mt} \geq 3$.

The S_{mt} multiplier is included in the calculation of MRC_{mt+1} for two reasons. First, as suggested by Hendricks and Hirtle (1997), it adjusts the reported VaR estimates up to what regulators consider to be a minimum capital requirement reflecting their concerns regarding prudent capital standards and model accuracy.[3] Second, S_{mt} explicitly links the accuracy of a bank's VaR model to its capital charge by varying over time. S_{mt} is set according to the accuracy of model m's VaR estimates for a one-day holding period ($k = 1$) and 99 percent coverage, denoted $\text{VaR}_{mt}(1,1)$ or simply VaR_{mt}.

S_{mt} is a step function that depends on the number of exceptions observed over the last 250 trading days. Exceptions are defined as occasions when the portfolio return ε_{t+1} is less than the corresponding VaR_{mt}.[4] The possible number of exceptions is divided into three zones. Within the green zone of four or fewer exceptions, a VaR model is deemed "acceptably accurate" to the regulators, and S_{mt} remains at its minimum value of three. Within the yellow zone of five to nine exceptions, S_{mt} increases incrementally with the number of exceptions. Within the red zone of ten or more exceptions, the VaR model is deemed to be "inaccurate" for regulatory purposes, and S_{mt} increases to its maximum value of four. The institution also must take explicit steps to improve its risk management system.[5]

The "internal models" approach represents a significant change in the regulatory oversight of bank trading activities, since previous approaches used relatively simple rules not based on bank-specific inputs. This approach indicates a move toward incentive-compatible regulations, i.e., regulations that give banks incentives to comply with desired outcomes.[6] Having established that market risk capital will be a function of banks' own VaR estimates, regulators must now focus on evaluating the accuracy of these VaR estimates. The following section discusses three methods for evaluating VaR estimates. In accordance with the current regulatory framework, one-step-ahead VaR estimates are analyzed.

Alternative Evaluation Methods

Under the MRA, regulators must determine whether a bank's VaR model is "acceptably accurate" given 250 VaR estimates and the corresponding portfolio returns. To date, two methods have been proposed for this type of evaluation: (1) evaluation based on the binomial distribution and (2) interval forecast evaluation, as proposed by Christoffersen (1998). Both methods use hypothesis tests to determine whether the VaR estimates exhibit a specified property that is characteristic of accurate VaR estimates.

However, as noted by Diebold and Lopez (1996), it is unlikely that forecasts from a model will exhibit all the properties of accurate forecasts. Thus, evaluating VaR estimates solely upon whether a specific property is present may yield only limited information regarding their accuracy. In addition, the power of the tests used in the evaluation must also be considered. In this paper, an evaluation method based on determining how well VaR estimates minimize a regulatory loss function is proposed. This evaluation method can provide information that is of direct interest to the regulators since their concerns are directly incorporated into the loss function.

Evaluation of VaR estimates based on the binomial distribution. Under the MRA, banks report their VaR estimates to the regulators, who observe when actual portfolio losses exceed these estimates.[7] As discussed by Kupiec (1995), assuming that the VaR estimates are accurate, such exceptions can be modeled as independent draws from a binomial distribution with a probability of occurrence equal to 1 percent. Accurate VaR estimates should exhibit the property that their unconditional coverage $\hat{\alpha} = x/250$, where x is the number of exceptions, equals 1 percent. Since the probability of observing x exceptions in a sample of size 250 under the null hypothesis is

3. See Stahl (1997) for a mathematical justification of the multiplication factor.

4. Note that the portfolio returns reported to the regulators, commonly referred to as the "profit & loss numbers," will usually not directly correspond to ε_{t+1}. The profit & loss numbers are usually polluted by the presence of customer fees and intraday trade results, which are not captured in standard VaR models. No definitive method of dealing with this discrepancy has been established.

5. The MRA contains a number of other criteria, such as "stress testing," that banks' risk management systems must meet in order to be considered appropriate for determining market risk capital requirements.

6. Note that an alternative (and possibly more incentive-compatible) method for monitoring the market risk exposure of commercial banks is the "precommitment" approach proposed by Kupiec and O'Brien (1995).

7. Note that this reporting is in dollar terms. Since the following discussion will be in terms of log portfolio returns, these reported numbers must be transformed into log form in order to make these evaluation methods operational.

$$\Pr(x) = \binom{250}{x} 0.01^x * 0.99^{250-x},$$

the appropriate likelihood ratio statistic for testing whether $\hat{\alpha} = 0.01$ is

$$LR_{uc} = 2[\log(\hat{\alpha}^x(1-\hat{\alpha})^{250-x}) - \log(0.01^x * 0.99^{250-x})].$$

Note that the LR_{uc} test is uniformly most powerful for a given sample size and that the statistic has an asymptotic $\chi^2(1)$ distribution.

The finite sample size and power characteristics of this test are of interest. With respect to size, the finite sample distribution of LR_{uc} for the specified parameters may be sufficiently different from a $\chi^2(1)$ distribution that the asymptotic critical values may be inappropriate. Table 1 Panel A presents the finite-sample critical values as determined via simulation and shows meaningful differences between the two distributions which must be accounted for when drawing statistical inference. As for the power of this test, Kupiec (1995) describes how this test has a limited ability to distinguish among alternative hypotheses and thus has low power in samples of size 250.

Evaluation of VaR estimates using the interval forecast method. VaR estimates are also interval forecasts of the lower 1 percent tail of f_{t+1}, the one-step-ahead return distribution. Interval forecasts can be evaluated conditionally or unconditionally, that is, with or without reference to the information available at each point in time. The LR_{uc} test is an unconditional test since it simply counts exceptions over the entire period. However, in the presence of time-dependent heteroskedasticity, the conditional accuracy of interval forecasts is an important issue. Interval forecasts that ignore such variance dynamics may have correct unconditional coverage but, at any given time, will have incorrect conditional coverage; see Figure 1 for an illustration. In such cases, the LR_{uc} test is of limited use since it will classify inaccurate VaR estimates as "acceptably accurate."

The LR_{cc} test, adapted from the more general test proposed by Christoffersen (1998), is a test of correct conditional coverage. For a given VaR estimate, the indicator variable I_{mt+1} for whether an exception occurred is constructed as

$$I_{mt+1} = \begin{cases} 1 & \text{if} \quad \varepsilon_{t+1} < \text{VaR}_{mt} \\ 0 & \text{if} \quad \varepsilon_{t+1} \geq \text{VaR}_{mt} \end{cases}.$$

Since accurate VaR estimates exhibit the property of correct conditional coverage, the I_{mt+1} series must exhibit both correct unconditional coverage and serial independence.

TABLE 1

CRITICAL VALUES FOR THE LR_{uc} AND LR_{cc} STATISTICS

	Significance Level		
	1%	5%	10%
	A. LR_{uc} Statistic		
Asymptotic $\chi^2(1)$	6.635	3.842	2.706
Finite-Sample	5.497 (0.5%)	5.025 (9.5%)	3.555 (12.2%)
	B. LR_{cc} Statistic		
Asymptotic $\chi^2(2)$	9.210	5.992	4.605
Finite-Sample	6.007 (0.2%)	5.015 (1.1%)	5.005 (11.8%)

NOTE: The finite-sample critical values for the LR_{uc} and LR_{cc} test statistics for the lower 1 percent quantile ($\alpha = 1$) are based on 10,000 simulations of sample size $T = 250$. The percentages in parentheses are the quantiles that correspond to the asymptotic critical values under the finite-sample distribution.

The LR_{cc} test is a joint test of these two properties. The relevant test statistic is $LR_{cc} = LR_{uc} + LR_{ind}$, which is asymptotically distributed $\chi^2(2)$. The finite sample critical values for the regulatory parameter values of $(k,\alpha) = (1,1)$ are shown in Table 1 Panel B.

The LR_{ind} statistic is the likelihood ratio statistic for the null hypothesis of serial independence against the alternative of first-order Markov dependence.[8] The likelihood function under this alternative hypothesis is

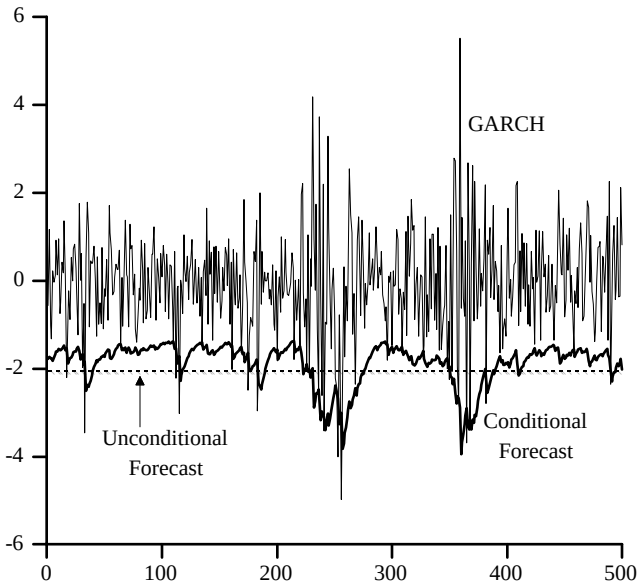
$$L_A = (1-\pi_{01})^{T_{00}} \pi_{01}^{T_{01}} (1-\pi_{11})^{T_{10}} \pi_{11}^{T_{11}},$$

where the T_{ij} notation denotes the number of observations in state j after having been in state i the period before, $\pi_{01} = T_{01}/(T_{00} + T_{01})$ and $\pi_{11} = T_{11}/(T_{10} + T_{11})$. Under the null hypothesis of independence, $\pi_{01} = \pi_{11} = \pi$, and the relevant likelihood function is $L_0 = (1-\pi)^{T_{00}+T_{10}} \pi^{T_{01}+T_{11}}$, where $\pi = (T_{01} + T_{11})/250$. The test statistic LR_{ind} is $2[\log L_A - \log L_0]$ and has an asymptotic $\chi^2(1)$ distribution.

8. As discussed in Christoffersen (1998), several other forms of dependence, such as second-order Markov dependence, can be specified. For the purposes of this paper, however, first-order Markov dependence is used.

FIGURE 1

GARCH(1,1)-NORMAL PROCESS WITH ONE-STEP-AHEAD, LOWER 5% CONDITIONAL AND UNCONDITIONAL INTERVAL FORECASTS



NOTE: The line labeled GARCH is a realization of 500 portfolio returns from a GARCH(1,1)-normal data-generating process. The variance dynamics are characterized as $h_{t+1} = 0.075 + 0.10\varepsilon_t^2 + 0.85h_t$, which imply an unconditional variance of 1.5. The unconditional interval forecasts are based on the unconditional $N(0,1^1/_2)$ distribution, and the conditional interval forecasts are based on the true data-generating process. Although both forecasts exhibit correct unconditional coverage with 25 exceptions (that is, $\alpha^* = \alpha = 5\%$), only the conditional confidence intervals exhibit correct conditional coverage or, in other words, provide 5% coverage at each point in time.

Evaluation of VaR estimates using regulatory loss functions. The loss function evaluation method proposed here is based not on a hypothesis-testing framework, but on assigning to VaR estimates a numerical score that reflects specific regulatory concerns. Although this method forgoes the benefits of statistical inference, it provides a measure of relative performance that can be used to compare VaR estimates across time and across institutions.

To use this method, the regulatory concerns of interest must be translated into a loss function. The general form of these loss functions is

$$C_{mt+1} = \begin{cases} f(\varepsilon_{t+1}, \text{VaR}_{mt}) & \text{if} \quad \varepsilon_{t+1} < \text{VaR}_{mt} \\ g(\varepsilon_{t+1}, \text{VaR}_{mt}) & \text{if} \quad \varepsilon_{t+1} \geq \text{VaR}_{mt} \end{cases},$$

where $f(x,y)$ and $g(x,y)$ are functions such that $f(x,y) \geq g(x,y)$. The numerical scores are constructed with a negative orientation; i.e., lower values of C_{mt+1} are preferred since exceptions are given higher scores than non-exceptions. Numerical scores are generated for individual VaR estimates, and the score for the complete regulatory sample is

$$C_m = \sum_{i=1}^{250} C_{mt+i} \, .$$

Under very general conditions, accurate VaR estimates will generate the lowest possible numerical score.[9] Once a loss function is defined and C_m is calculated, a benchmark can be constructed and used to evaluate the performance of a set of VaR_{mt} estimates. Although many regulatory loss functions can be constructed, the three analyzed in this paper are described below.

Loss function implied by the binomial method. The loss function implied by the binomial method is

$$C_{mt+1} = \begin{cases} 1 & \text{if} \quad \varepsilon_{t+1} < \text{VaR}_{mt} \\ 0 & \text{if} \quad \varepsilon_{t+1} \geq \text{VaR}_{mt} \end{cases}.$$

Note that the appropriate benchmark is $E[C_{mt+1}] = 0.01$, which for the full sample is $E[C_m] = 2.5$. As before, only the number of exceptions is of interest, and no additional information beyond that contained in the binomial method is included in this analysis.

Loss function analogous to the adjustment schedule for the S_{mt} multiplier. The numerical score assigned to a set of 250 VaR estimates can be generated by assigning a score to each element of the set or by assigning a score based on the entire set. The adjustment to the S_{mt} multiplier embodied in the MRA is based on the entire set of VaR estimates. Phrased in the notation above, the loss function that generates an analogous numerical score is

$$C_{mt+1}(x) = \begin{cases} 0 & \text{if} \quad \varepsilon_{t+1} \geq \text{VaR}_{mt} \\ 0 & \text{if} \quad \varepsilon_{t+1} < \text{VaR}_{mt} \quad \text{and} \quad 0 < x \leq 4 \\ 0.4/5 & \text{if} \quad \varepsilon_{t+1} < \text{VaR}_{mt} \quad \text{and} \quad x = 5 \\ 0.5/6 & \text{if} \quad \varepsilon_{t+1} < \text{VaR}_{mt} \quad \text{and} \quad x = 6 \\ 0.65/7 & \text{if} \quad \varepsilon_{t+1} < \text{VaR}_{mt} \quad \text{and} \quad x = 7 \\ 0.75/8 & \text{if} \quad \varepsilon_{t+1} < \text{VaR}_{mt} \quad \text{and} \quad x = 8 \\ 0.85/9 & \text{if} \quad \varepsilon_{t+1} < \text{VaR}_{mt} \quad \text{and} \quad x = 9 \\ 1/x & \text{if} \quad \varepsilon_{t+1} < \text{VaR}_{mt} \quad \text{and} \quad x \geq 10 \end{cases},$$

9. See Diebold, Gunther, and Tay (1998) as well as Granger and Pesaran (1996) for further discussion of these conditions with respect to distribution and probability forecasts, respectively.

where x is the number of exceptions in the entire sample and the numerical weights are the actual S_{mt} values divided by x.[10] The benchmark for this numerical score is

$$E\left[C_m(x)\right] = E\left[\sum_{i=1}^{250} C_{mt+i}(x)\right] = \sum_{x=0}^{250} \Pr(x) * \left(\sum_{i=1}^{250} C_{mt+i}(x) | x\right) = 0.05597 .$$

Note that this loss function incorporates the regulatory concerns expressed in the S_{mt} multiplier, but, like the binomial loss function, it is based only on the number of exceptions in the sample.

Loss function that addresses the magnitude of the exceptions. As noted by the Basle Committee on Banking Supervision (1996), the magnitude as well as the number of the exceptions are a matter of concern to regulators.[11] As discussed by Hendricks (1996), the magnitude of the observed exceptions can be quite large; in that study, the portfolio losses exceed the corresponding VaR estimate by 30 to 40 percent on average, and, in the extreme cases, by up to 300 percent.

This concern can readily be incorporated into a loss function by introducing a magnitude term into the binomial loss function. Although several are possible, a quadratic term is used here, such that

$$C_{mt+1} = \begin{cases} 1 + (\varepsilon_{t+1} - \text{VaR}_{mt})^2 & \text{if} \quad \varepsilon_{t+1} < \text{VaR}_{mt} \\ 0 & \text{if} \quad \varepsilon_{t+1} \geq \text{VaR}_{mt} \end{cases} .$$

Thus, as before, a score of 1 is imposed when an exception occurs, but now an additional term based on the magnitude of the exception is included. The numerical score increases with the magnitude of the exception and can provide additional information on how the underlying VaR model forecasts the lower tail of the f_{t+1} distribution. Unfortunately, the benchmark based on the expected value of C_{mt+1} cannot easily be determined because the f_{t+1} distribution is unknown. However, simple operational benchmarks based on certain distributional assumptions can be constructed and are discussed in Section III.

10. As currently constructed, the S_{mt} adjustment schedule does not address VaR estimates that are possibly too conservative, i.e., VaR estimates that lead to a lower than expected number of exceptions. Given the regulatory interest in providing adequate capital against negative outcomes, the absence of such outcomes is not relevant. However, from the perspective of VaR model evaluation, such outcomes might indicate modeling error. This concern could be addressed by modifying the loss function to include a non-zero score when $x < 4$.

11. Note that Berkowitz (1999) has recently developed a hypothesis-testing method for calculating VaR estimates that incorporates the magnitudes of the exception.

II. SIMULATION EXERCISE

To analyze the ability of the three evaluation methods to gauge the accuracy of VaR estimates and thus avoid VaR model misclassification, a simulation exercise is conducted. For the two hypothesis-testing methods, this amounts to analyzing the power of the statistical tests, i.e., determining the probability with which the tests reject the specified null hypothesis when it is incorrect. With respect to the loss function method, its ability to evaluate VaR estimates is gauged by how frequently the numerical score for VaR estimates generated from the true data-generating process (DGP) is lower than the score for the VaR estimates from alternative models. If this method is capable of distinguishing between these competing scores, then the degree of VaR model misclassification will be low.

The first step in this simulation exercise is deciding what type of portfolio to analyze. Although VaR models are commonly applied to complicated portfolios of financial assets, the log portfolio value y_{t+1} used here is specified as $y_{t+1} = y_t + \varepsilon_{t+1}$, where $\varepsilon_{t+1} \mid \Omega_t \sim f_{t+1}$. This process is representative of linear deterministic conditional mean specifications. It is only for portfolios with nonlinear elements, such as portfolios with options, that this choice presents inference problems; further research along these lines, as by Pritsker (1997), is needed.

The simulation exercise is conducted in four parts. To examine how the evaluation methods perform under different distributional assumptions, the true DGP f_{t+1} is set to be the standard normal distribution and a t-distribution with six degrees of freedom, which induces fatter tails than the normal, in the first two parts. The last two parts examine the performance of the evaluation methods in the presence of variance dynamics: the third part models ε_{t+1} as a GARCH(1,1)-normal process, and the fourth part does so as a GARCH(1,1)-t(6) process.

In each part of the exercise, the true DGP is one of eight VaR models evaluated and is designated as the true model or model 1. Traditional power analysis of a hypothesis test is conducted by varying a particular parameter and determining whether the corresponding incorrect null hypothesis is rejected; such changes in parameters generate what are termed local alternatives. In this study, non-nested, but common, VaR models are used as reasonable "local" alternatives. For example, a common type of VaR model specifies the variance of ε_{t+1}, denoted as h_{mt+1}, as an exponentially weighted moving average of squared innovations; that is,

$$h_{mt+1} = (1-\lambda)\sum_{i=0}^{\infty} \lambda^i \varepsilon_{t-i}^2 = \lambda h_{mt} + (1-\lambda)\varepsilon_t^2 .$$

This VaR model, a version of which is used in the well-known RiskMetrics calculations (see J.P. Morgan, 1996), is calibrated here by setting λ equal to 0.94 or 0.99, which imply a high-degree of persistence in variance.[12] The alternative VaR models used in each part of the simulation exercise are described in the subsections below.

The simulation runs are structured identically in each part of the exercise. For each run, the simulated y_{t+1} series is generated using the chosen DGP. After generating an in-sample period of 3,500 observations, the chosen VaR models are used to generate one-step-ahead VaR estimates for the next 250 out-of-sample observations of y_{t+1}. The analytical results are based on 1,000 simulation runs. The simulation results are organized below with respect to the four parts of the exercise.

Two general points can be made regarding the simulation results. First, with the size of the tests set at 5 percent, the power of the two hypothesis-testing methods varies considerably against the incorrect null hypotheses implied by the alternative VaR models. In some cases, the power of the tests is high (greater than 75 percent), but in the majority of the cases examined, the power is poor (less than 50 percent) to moderate (between 50 and 75 percent). The results indicate that these two methods are likely to misclassify VaR estimates from inaccurate models as "acceptably accurate."

Second, the degree of model misclassification exhibited by the loss function method roughly matches that of the other two methods; i.e., when the hypothesis-testing methods exhibit low power, the loss function method is also generally less capable of distinguishing between accurate and inaccurate VaR estimates. Overall, however, the loss function method has a moderate to high ability to gauge the accuracy of VaR estimates. Among the three regulatory loss functions, the results for the magnitude loss function are relatively better, indicating a greater ability to distinguish between models. This result is not surprising given that the magnitude loss function incorporates additional information—the magnitude of the exceptions—into the evaluation.

Simulation Results for the Homoskedastic Standard Normal DGP

For the first part of this exercise, the true DGP is the standard normal; i.e., $\varepsilon_{t+1} \sim N(0,1)$. The seven alternative models examined are: normal distributions with variances of $\frac{1}{2}$, $\frac{3}{4}$, $1\frac{1}{4}$, and $1\frac{1}{2}$; the two heteroskedastic calibrated VaR models with normal distributions; and the historical simulation model. For this last model, the VaR estimates are formed as the lower 1 percent quantile of the empirical distribution of the 500 previously observed returns.

In Table 2, Panel A presents the power analysis of the hypothesis-testing evaluation methods for a fixed test size of 5 percent. For the homoskedastic alternatives (models 2 through 5), the power results vary considerably and are related to the differences between the true variance and the model's variance; i.e., larger differences lead to greater relative power. With respect to the calibrated models (6 and 7), the tests have no power since the VaR estimates are still quite similar to those of the true DGP, even though unnecessary heteroskedasticity is introduced. Both tests have low power for the historical simulation model (model 8).

Panel B contains the comparative accuracy results for the loss function method using the specified loss functions. Note that each number in the panel represents the percentage of simulations for which the numerical score for the true model is actually lower than that of the inaccurate model. This method cannot distinguish between the numerical scores for the true DGP and those for models 4 and 5, which generate conservative VaR estimates and thus lower scores due to fewer exceptions. This result is generally acceptable from the viewpoint implicit in the regulatory loss functions, since regulators are concerned if not enough capital is held against possible losses, but not if too much capital is held. However, this method clearly can distinguish between the true DGP and the low variance models (models 2 and 3) that consistently generate smaller VaR estimates than necessary. With respect to the calibrated and historical models (models 6 through 8), the degree of misclassification is generally moderate, although the magnitude loss function exhibits the best results.

Simulation Results for the Homoskedastic t(6) DGP

For the second part of the exercise, the true DGP is a $t(6)$ distribution; i.e., $\varepsilon_{t+1} \sim t(6)$. The seven alternative models are: two normal distributions with variances of 1 and $1\frac{1}{2}$ (the same variance as the true DGP); the two calibrated models with normal distributions as well as with $t(6)$ distributions; and the historical simulation model.

12. Note that this VaR model is often implemented with a finite lag-order. For example, the infinite sum is frequently truncated at 250 observations, which accounts for over 90 percent of the sum of the weights. See Hendricks (1996) for further discussion on the choice of λ and the truncation lag. In this simulation exercise, no such truncation is imposed but, of course, one is implied by the overall sample size of the simulated time series.

TABLE 2

SIMULATION RESULTS FOR HOMOSKEDASTIC STANDARD NORMAL DGP

	Models						
	Homoskedastic				Heteroskedastic		Historical
	2	3	4	5	6	7	8
	A. Power of the LR_{uc} and LR_{cc} Tests against Alternative VaR Models[a] (%)						
LR_{uc}	97.2	30.4	29.7	54.9	4.3	4.5	40.2
LR_{cc}	97.8	32.9	30.5	60.1	5.4	5.7	43.4
	B. Accuracy of VaR Estimates Using Regulatory Loss Functions[b] (%)						
Binomial	100.0	94.4	0.0	0.0	55.3	55.4	28.3
Zone	99.6	66.8	0.0	0.0	17.9	18.2	6.7
Magnitude	100.0	99.7	0.0	0.0	76.1	76.4	53.8

[a] The size of the tests is set at 5% using the finite-sample critical values in Table 1.

[b] Each row represents the percentage of simulations for which the alternative VaR estimates have a higher numerical score than the true model, i.e., the percentage of the simulations for which the alternative VaR estimates are correctly classified as inaccurate.

NOTE: The results are based on 1,000 simulations. Model 1 is the true data generating process, $\varepsilon_{t+1} \sim N(0,1)$. Models 2 through 5 are homoskedastic normal distributions with variances of $1/2$, $3/4$, $1\,1/4$, and $1\,1/2$, respectively. Models 6 and 7 are normal distributions whose variances are exponentially weighted averages of the squared innovations calibrated using $\lambda = 0.94$ and $\lambda = 0.99$, respectively. Model 8 is the historical simulation model based on the previous 500 observations.

In Table 3, Panel A shows that the overall power of the *LR* tests against these alternative models is low. With the exception of the $N(0,1)$ model (model 2), the power results are below 50 percent; thus, the alternative VaR estimates are incorrectly classified as "acceptably accurate" a large percentage of the time. This result is mainly due to the similarity of the alternative VaR models to the true DGP. For example, although models 4 through 7 introduce unnecessary heteroskedasticity, their VaR estimates are similar to the true, but constant, VaR estimates.

Panel B contains the results of the loss function evaluation. For the normality-based models (models 2 through 5), the three loss functions have moderate to high ability to distinguish between alternative VaR estimates, with the zone loss function doing worst and the magnitude loss function doing best. However, with respect to models 6 through 8, this method shows a high degree of model misclassification due to the models' similarity to the true DGP.

Simulation Results for the GARCH(1,1)-normal DGP

For the last two parts of the simulation exercise, variance dynamics are introduced by using conditional heteroskedasticity of the GARCH form; i.e., $h_{t+1} = 0.075 + 0.10\varepsilon_t^2 + 0.85h_t$, which has an unconditional variance of $1\,1/2$. The only difference between the DGPs in these two parts of the exercise is the chosen distributional form. For the third part, $\varepsilon_{t+1} \mid \Omega_t \sim N(0, h_{t+1})$, and for the fourth part, $\varepsilon_{t+1} \mid \Omega_t \sim t(h_{t+1}, 6)$. The seven alternative VaR models examined are the homoskedastic models of the standard normal, the $N(0,1\,1/2)$ model, and the $t(6)$ distribution; the historical simulation model; and the calibrated models with normal innovations and the GARCH model with the other distributional form.

In Table 4, Panel A presents the power analysis of the hypothesis-testing methods. The power results are mainly driven by the differences between the distributional assumptions used by the true DGP and the alternative models. Specifically, the tests have low power against the calibrated normal models (models 5 and 6) since their smoothed variances are quite similar to the true GARCH variances. However, the results for the GARCH-$t(6)$ model (model 7) are much better due to the incorrect $t(6)$ assumption. Overall, the hypothesis-testing methods seem to have substantially less power against VaR models characterized by close approximations of the true variance dynamics (models 3 through 6 and 8) and have better power against models with incorrect distributional assumptions (models 2 and 7).

The results for the loss function evaluation method, presented in Panel B, are similar; that is, this method has a low to moderate ability to distinguish between the true and alternative VaR models. For the heteroskedastic models,

TABLE 3

SIMULATION RESULTS FOR HOMOSKEDASTIC $t(6)$ DGP

	Models						
	Homoskedastic		Heteroskedastic				Historical
	2	3	4	5	6	7	8
	A. POWER OF THE LR_{uc} AND LR_{cc} AGAINST ALTERNATIVE VAR MODELS[a] (%)						
LR_{uc}	59.1	10.8	15.3	14.6	20.3	19.9	7.9
LR_{cc}	61.5	11.2	17.4	19.9	30.4	30.5	12.4
	B. ACCURACY OF VAR MODELS USING REGULATORY LOSS FUNCTIONS[b] (%)						
Binomial	99.2	69.8	85.5	85.5	5.1	5.0	26.3
Zone	85.0	27.1	47.5	47.3	0.2	0.1	5.4
Magnitude	99.9	97.4	97.3	97.2	10.7	10.3	51.0

[a] The size of the tests is set at 5% using the finite-sample critical values in Table 1.

[b] Each row represents the percentage of simulations for which the alternative VaR estimates have a higher numerical score than the true model, i.e., the percentage of the simulations for which the alternative VaR estimates are correctly classified as inaccurate.

NOTE: The results are based on 1,000 simulations. Model 1 is the true data generating process, $\varepsilon_{t+1} \sim t(6)$. Models 2 and 3 are the homoskedastic models with normal distributions of variance of 1 and 1.5, respectively. Models 4 and 5 are the calibrated heteroskedastic models with the normal distribution, and models 6 and 7 are the calibrated heteroskedastic models with the $t(6)$ distribution. Model 8 is the historical simulation model based on the previous 500 observations.

TABLE 4

SIMULATION RESULTS FOR GARCH(1,1)-NORMAL DGP

	Models						
	Homoskedastic			Heteroskedastic			Historical
	2	3	4	5	6	7	8
	A. POWER OF THE LR_{uc} AND LR_{cc} AGAINST ALTERNATIVE VAR MODELS[a] (%)						
LR_{uc}	52.3	21.4	30.5	5.1	10.3	81.7	23.2
LR_{cc}	56.3	25.4	38.4	6.7	11.9	91.6	33.1
	B. ACCURACY OF VAR MODELS USING REGULATORY LOSS FUNCTIONS[b] (%)						
Binomial	91.7	41.3	18.1	52.2	48.9	0.0	38.0
Zone	72.1	21.0	8.1	15.2	18.4	0.0	17.7
Magnitude	96.5	56.1	29.1	75.3	69.4	0.0	51.5

[a] The size of the tests is set at 5% using the finite-sample critical values in Table 1.

[b] Each row represents the percentage of simulations for which the alternative VaR estimates have a higher numerical score than the true model, i.e., the percentage of the simulations for which the alternative VaR estimates are correctly classified as inaccurate.

NOTE: The results are based on 1,000 simulations. Model 1 is the true data generating process, $\varepsilon_{t+1} \mid \Omega_t \sim N(0,h_{t+1})$. Models 2, 3, and 4 are the homoskedastic models $N(0, 1)$, $N(0,1.5)$ and $t(6)$, respectively. Models 5 and 6 are the two calibrated heteroskedastic models with the normal distribution, and model 7 is a GARCH(1,1)-$t(6)$ model with the same parameter values as Model 1. Model 8 is the historical simulation model based on the previous 500 observations.

the more conservative GARCH-*t*(6) model (model 7) obviously minimizes the loss functions due to its smaller number of exceptions. For the calibrated normal models (models 5 and 6) and the historical model (model 8), this method generally has a poor ability to classify them correctly. With respect to the homoskedastic models (models 2 through 4), the degree of misclassification is low for the standard normal (model 2), but much higher for the other two models that have the same unconditional variance as the true DGP. Note, as previously mentioned, that the magnitude loss function is relatively more able to classify VaR estimates correctly than the other two loss functions.

Simulation Results for the GARCH(1,1)-t(6) DGP

In Table 5, Panel A presents the power analysis of the hypothesis-testing methods. The power results are clearly tied to the presence of heteroskedasticity in the alternative VaR models. The homoskedastic models (models 2 through 4) are identified as "inaccurate" with very high power since their VaR estimates cannot match the magnitude of the observed returns from the true DGP. However, for the heteroskedastic models (models 5 through 7) and the historical model (model 8), which are more capable of tracking the underlying variance, the power of the tests declines dramatically.

For the loss function method, the results in Panel B indicate that the VaR estimates from the true and alternative models, except for the historical model (model 8), can be differentiated. For the homoskedastic alternatives (models 2 through 4), this ability is driven mainly by the fact that constant VaR estimates cannot track the actual returns process well. The heteroskedastic models (models 5 though 7) that can adjust over time do better, but they can still be identified as inaccurate due to their misspecified distributional assumptions. For the historical model, the method's ability to distinguish it from the true DGP is diminished. Note again that, of the three loss functions, the magnitude loss function is most capable of differentiating between the models.

III. IMPLEMENTATION OF THE LOSS FUNCTION METHOD

The simulation results presented above indicate that the loss function method is generally capable of distinguishing between VaR estimates from the true DGP and alternative models. Although this ability varies, the method can provide information useful for the regulatory evaluation of VaR estimates, particularly when the magnitude loss function is used. This result is not surprising given that it incorporates the additional information on the magnitude of the

TABLE 5

SIMULATION RESULTS FOR GARCH(1,1)-*t*(6) DGP

	MODELS						
	Homoskedastic			Heteroskedastic			Historical
	2	3	4	5	6	7	8
	A. POWER OF THE LR_{uc} AND LR_{cc} AGAINST ALTERNATIVE VAR MODELS[a] (%)						
LR_{uc}	99.8	97.5	94.4	17.9	34.7	59.1	47.3
LR_{cc}	99.9	97.7	95.6	23.7	35.6	61.5	54.8
	B. ACCURACY OF VAR ESTIMATES USING REGULATORY LOSS FUNCTIONS[b] (%)						
Binomial	99.9	99.9	99.8	82.6	66.9	99.2	42.4
Zone	99.9	99.0	97.1	47.2	42.7	85.0	29.9
Magnitude	99.9	99.9	99.9	94.8	78.0	99.9	53.7

[a] The size of the tests is set at 5% using the finite-sample critical values in Table 1.

[b] Each row represents the percentage of simulations for which the alternative VaR estimates have a higher numerical score than the true model, i.e., the percentage of the simulations for which the alternative VaR estimates are correctly classified as inaccurate.

NOTE: The results are based on 1,000 simulations. Model 1 is the true data generating process, $\varepsilon_{t+1} \mid \Omega_t \sim t(h_{t+1},6)$. Models 2, 3, and 4 are the homoskedastic models $N(0,1)$, $N(0,1.5)$ and $t(6)$, respectively. Models 5 and 6 are the two calibrated heteroskedastic models with the normal distribution, and model 7 is a GARCH(1,1)-normal model with the same parameter values as Model 1. Model 8 is the historical simulation model based on the previous 500 observations.

exceptions into the evaluation. In this section, this evaluation method using the magnitude loss function is made operational by creating a benchmarking process and by illustrating its use in a detailed example.

Creating a Benchmark for the Observed Numerical Scores

Under the current regulatory framework, regulators observe the VaR estimates and portfolio returns, denoted $\{\varepsilon_{t+i}, \text{VaR}_{mt+i-1}\}_{i=1}^{250}$, for bank m and thus can construct, under the magnitude loss function, the numerical score C_m. However, for a particular realized value C_m^*, aside from the number of exceptions, not much inference on the performance of the underlying VaR estimates is available. That is, we don't know whether C_m^* is a "high" or "low" number. Although comparisons could be made cross-sectionally across banks, a better method for gauging the magnitude of C_m^* is to create a comparative benchmark based on the distribution of C_m, which is a random variable due to the random portfolio returns. Since each portfolio return has a conditional distribution $\varepsilon_{t+1} \mid \Omega_t \sim f_{t+1}$, additional assumptions on the dependence of the returns and their distributions must be imposed in order to analyze $f(C_m)$, the distribution of C_m.

An immediate and commonly used assumption is that the observed returns are independent and identically distributed (iid); i.e., $\varepsilon_{t+1} \sim f$. This is quite a strong assumption, especially given the heteroskedasticity often found in portfolio returns.[13] However, the small sample size of 250 observations mandated by the MRA allows few other choices. Having made the assumption that the observed returns are iid, their empirical distribution, denoted $\hat{f}(\varepsilon_{t+1})$, can be estimated using a variety of methods. For example, nonparametric methods, such as smoothed kernel density estimators as per Silverman (1986) or unsmoothed bootstrap methods, could be used. Generally, for issues of tractability, parametric methods are commonly used; i.e., a specific distributional form is assumed, and the necessary parameters are estimated from the available data. For example, if the returns are assumed to be normally distributed with zero mean, the variance can be estimated such that $\hat{f}(\varepsilon_{t+1})$ is $N(0,\hat{\sigma}^2)$.

A reasonable alternative to assuming independence is to impose some explicit form of dependence on the data. For example, if the returns are assumed to be driven by a GARCH process, the necessary parameters could be estimated from the observed portfolio returns and used to specify $\hat{f}(\varepsilon_{t+1} \mid \Omega_t)$. Since the small sample size will limit the usefulness of such parameter estimates, the calibrated models previously discussed present a reasonable alternative specification.[14] In the example that follows, both assumptions are used to examine the VaR estimates for different models.

Once $\hat{f}(\varepsilon_{t+1})$ or $\hat{f}(\varepsilon_{t+1} \mid \Omega_t)$ has been determined, the empirical distribution of the numerical score C_m under the distributional assumptions, denoted $\hat{f}(C_m)$, can be generated. For example, if $\varepsilon_{t+1} \sim N(0,\hat{\sigma}^2)$, then the corresponding VaR estimates are $\text{VaR}_{\hat{f}t} = -2.32\hat{\sigma}$. If the assumption is that $\varepsilon_{t+1} \sim N(0,\hat{h}_{t+1})$, then

$$\text{VaR}_{\hat{f}t} = -2.32\sqrt{\hat{h}_{t+1}},$$

where $\hat{h}_{t+1}$ is the assumed variance at time $t+1$. Using these assumptions, $\hat{f}(C_m)$ can then be constructed via simulation by forming, say, 1,000 values of the numerical score C_m, each based on 250 draws from the assumed distribution of ε_{t+1} and its corresponding VaR estimates.[15]

Once $\hat{f}(C_m)$ has been generated, the empirical quantile $\hat{q}_m^* = \hat{F}(C_m^*)$, where $\hat{F}(C_m)$ is the cumulative distribution function of $\hat{f}(C_m)$, can be calculated for the observed value C_m^*. This empirical quantile provides a performance benchmark, based on the distributional assumptions, that can be incorporated into the regulatory evaluation of the underlying VaR estimates. In order to make this benchmark operational, the regulator should select a threshold quantile above which concerns regarding the performance of the VaR estimates are raised. This decision should be based both on the regulators' preferences and the severity of the distributional assumptions used. If $\hat{q}_m^*$ is below the threshold that regulators believe is appropriate, say, below 80 percent, then C_m^* is "typical" under the assumptions made about the portfolio returns and given the regulators' preferences. If $\hat{q}_m^*$ is above the threshold, then C_m^* can be considered atypical given their preferences, and the regulators should take a closer look at the underlying VaR model.

Note that this method for evaluating VaR estimates does not replace the hypothesis-testing methods but, instead, provides complementary information, especially regarding the magnitude of the exceptions. In addition, the flexibility of this method permits many other concerns to be incorporated into the analysis via the choice of the loss function.

13. See Kearns and Pagan (1997) for a discussion of the consequences of ignoring the dependence in financial data when drawing inferences about the tails of the data's distribution.

14. The negative impact of misspecified dependence in the data on the construction of $\hat{f}(C_m)$ relative to that of the iid assumption is not known; further research is necessary.

15. Note that although a closed form solution for $\hat{f}(C_m)$ should be available if a parametric assumption is made, simulation methods will be used in this paper.

The example below illustrates how this method might be employed in an actual case; it can be seen that, in certain cases, the loss function method flags important information not captured in the standard binomial analysis.

Detailed Example

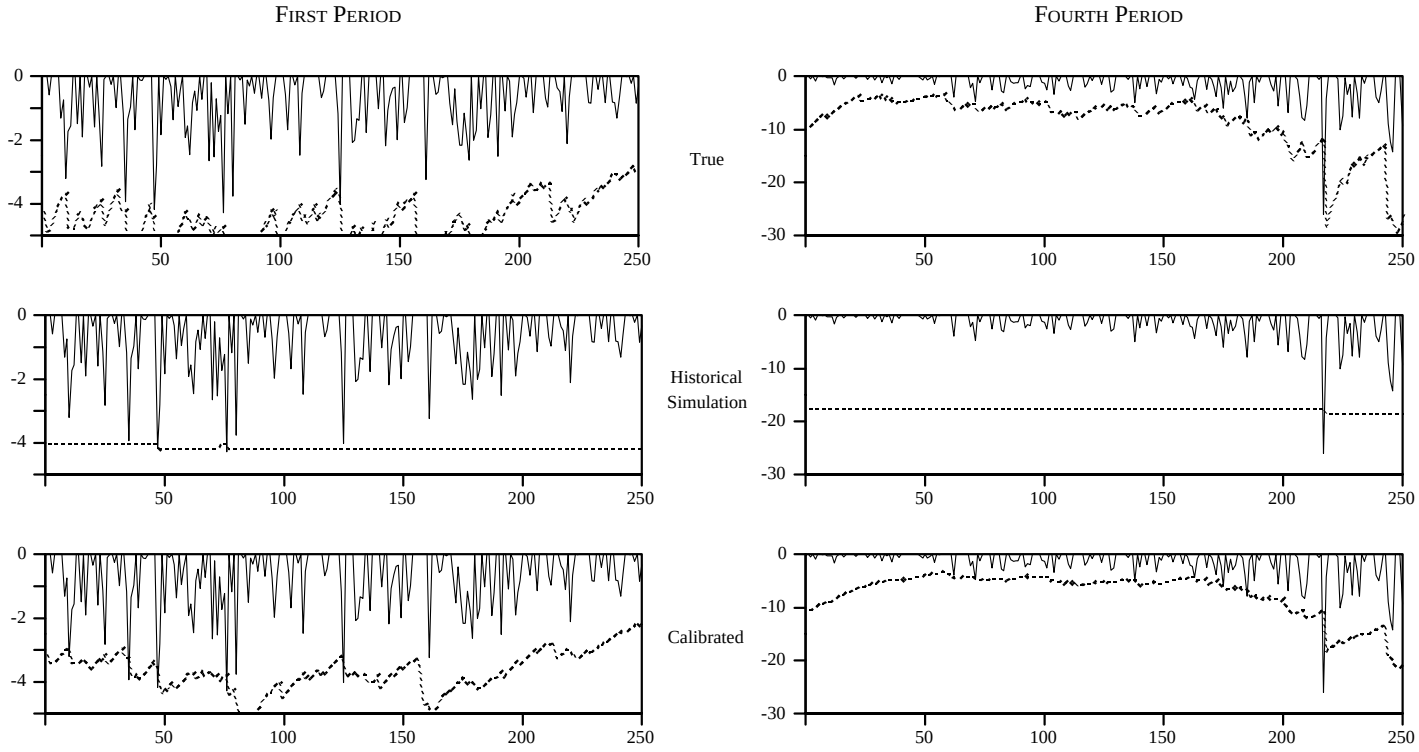
For this detailed example, the performance of three sets of VaR estimates is examined using the three evaluation methods. As will be shown, inferences about the accuracy of the VaR estimates based on the loss function method match those drawn from the hypothesis-testing methods. However, since it incorporates additional information on the magnitude of the exceptions, the loss function method permits regulators to draw further inferences.

The underlying returns process is $\varepsilon_{t+1} \mid \Omega_t \sim t(h_{t+1},6)$ with $h_{t+1} = 0.075 + 0.10\varepsilon_t^2 + 0.85h_t$. VaR estimates are generated from three VaR models: the true GARCH-*t*(6) model; the historical simulation model based on a rolling window of the 500 previous observations; and the calibrated normal model with $\lambda = 0.94$. The models are henceforth denoted as the true, historical, and calibrated models, respectively. The 1,250 generated observations are analyzed over the five contiguous but non-overlapping periods of 250 observations. Two periods of simulated data and the corresponding VaR estimates are plotted in Figure 2.

Table 6 contains the evaluation results for the two hypothesis-testing methods. Panel A reports the number of exceptions in each of the five periods for the three sets of VaR estimates, and Panels B and C report the LR_{uc} and LR_{cc} statistics, respectively. The occasions for which the null hypothesis is rejected at the 5 percent significance level are noted. For the true model, both tests correctly do not reject the null hypothesis that the VaR estimates exhibit the specified properties, and the S_{mt} multiplier would remain at three. For the historical model, the number of exceptions is particularly large in the second and third periods, and the corresponding test statistics reject the null hypotheses. In

FIGURE 2

SIMULATED SERIES FOR THE ILLUSTRATION



NOTE: The solid line represents the simulated negative returns. The dotted line represents the corresponding VaR estimates from each of the three models. The points at which the solid line crosses the dotted line are the exceptions in the sample. Note that, for the true and calibrated models, VaR estimates that are more negative than permitted by the specified y-axis are not shown.

these periods, S_{mt} increases to its maximum value of four. However, for the other time periods, this VaR model is "acceptably accurate," even though the hypothesis tests indicate a problem in the fifth period when no exceptions occurred. For the calibrated model, the null hypotheses are rejected in only one case, and S_{mt} is above three in all but one period. These results present a tangible example of the poor power characteristics of these tests.

Turning to the proposed loss function method, in Table 7, Panel A contains the C_m^* numerical scores under the magnitude loss function. As mentioned, these scores alone do not provide a very useful basis for evaluating the VaR estimates.

TABLE 6

HYPOTHESIS-TESTING RESULTS FOR THE DETAILED EXAMPLE

	Model		
Period	True	Historical	Calibrated
	A. Number of Exceptions		
1	1	2	5
2	3	11	6
3	1	14	3
4	1	1	6
5	2	0	5
	B. LR_{uc} Statistics		
1	1.1765	0.1084	1.9568
2	0.0949	15.8906*	3.5554
3	1.1765	25.7803*	0.0949
4	1.1765	1.1765	3.5554
5	0.1084	5.0252*	1.9568
	C. LR_{cc} Statistics		
1	1.1846	0.1408	2.1617
2	0.1681	16.9078*	3.8517
3	1.1846	30.1907*	5.5202*
4	1.1846	1.1846	3.8517
5	0.1408	5.0252*	2.1617

NOTE: The time periods are based on a division of the entire simulation run of 1,250 observations into five contiguous, but non-overlapping periods of 250 observations. The true model is $\varepsilon_{t+1} \mid \Omega_t \sim t(h_{t+1},6)$ with $h_{t+1} = 0.075 + 0.10\varepsilon_t^2 + 0.85h_t$. The historical simulation model is based on the 500 previous observations. The calibrated model uses the calibrated variance parameter of $\lambda = 0.94$ and the normal distribution.

The asterisk indicates that the null hypothesis is rejected at the 5% significance level using the finite-sample critical values presented in Table 1.

TABLE 7

MAGNITUDE LOSS FUNCTION RESULTS FOR THE DETAILED EXAMPLE

	Model		
Period	True	Historical	Calibrated
	A. Numerical Scores		
1	1.1287	2.0803	7.1048
2	3.8180	58.3150	15.8955
3	1.4854	507.5814	24.7188
4	200.1094	71.4351	243.8740
5	15.6136	0.0	16.9524
	B. Empirical Quantiles under the True DGP (%)		
1	13.7	22.0	54.1
2	31.0	**89.6**	64.6
3	11.3	**86.9**	37.6
4	**95.9**	**86.1**	**97.0**
5	53.8	0.0	56.1
	C. Empirical Quantiles under the Normal Distribution (%)		
1	17.4	29.0	**88.6**
2	44.0	**100.0**	**91.7**
3	10.5	**99.8**	46.5
4	**100.0**	**99.8**	**100.0**
5	**82.7**	0.0	**84.3**
	D. Empirical Quantiles under the Calibrated Normal Distribution (%)		
1	20.5	33.1	**90.1**
2	52.9	**99.6**	**93.4**
3	13.6	**99.3**	62.2
4	**99.9**	**98.8**	**99.9**
5	**86.7**	0.0	**88.5**

NOTE: See note to Table 6.

However, by making assumptions about the distribution of the observed returns, an approximate distribution of the numerical scores, $\hat{f}(C_m)$, can be generated via simulation and used to provide a benchmark for evaluation.

Since, in this example, the true DGP is known, the actual $f(C_m)$ can be generated. In Table 7, Panel B reports the empirical quantiles q_m^* under $f(C_m)$. Three results are immediately clear. First, the inference drawn from the loss function method generally matches that drawn from the two hypothesis-testing methods; i.e., the q_m^*s are generally low (below the threshold 80 percent), except in a few distinct cases. Second, the q_m^*s for the historical model in the second and third periods are high (above 80 percent) due to the large number of exceptions. Third, the q_m^*s for all of the models are high in the fourth period, even though the number of exceptions is low. Recall that the two hypothesis-testing methods indicated that these three sets of VaR estimates were "acceptably accurate" for that period.

The reason for the relatively high scores and q_m^*s in the fourth period can be seen in Figure 2. Observation number 217 is a particularly large negative number; in terms of relative magnitudes, it exceeds the VaR estimates by about 120 percent for the true model, 50 percent for the historical model and 144 percent for the calibrated model. This result clearly indicates the advantages of the loss function evaluation. By incorporating additional information on the magnitude of the exceptions into the evaluation, this method can alert the regulator when an extraordinary event, not detectable by the hypothesis-testing methods, has occurred.

In an actual implementation of the loss function evaluation method, the true DGP is not known. Hence, Panels C and D of Table 7 contain the q_m^*s under two different assumed $\hat{f}(C_m)$ distributions. In Panel C, $\hat{f}(C_m)$ is formed under the assumption that the returns are independent and normally distributed; i.e., $\varepsilon_{t+1} \sim N(0,\hat{\sigma}^2)$. In Panel D, $\hat{f}(C_m \mid \Omega_t)$ is formed under the assumption that $\varepsilon_{t+1} \sim N(0,\hat{h}_{t+1})$, where $\hat{h}_{t+1}$ follows an exponentially weighted moving average of squared observed returns with a calibration parameter of 0.94.[16] The empirical quantiles under these two assumed distributions are higher than those under the true DGP, which causes a form of Type I error; that is, under these assumed distributions and for a fixed threshold quantile, the observed C_m^*s will indicate more instances of possibly large exceptions than are called for under the true DGP. The reason for this upward bias is that under these distributional assumptions, the expected value of C_{mt+1} conditional on an exception having occurred will be lower than under the true DGP, skewing $\hat{f}(C_m)$ and $\hat{f}(C_m \mid \Omega_t)$ more towards zero than the true $f(C_m)$ distribution.[17] Thus, when the C_m^*s are compared to $\hat{f}(C_m)$ and $\hat{f}(C_m \mid \Omega_t)$, they will generally be in a higher quantile than under $f(C_m)$.

Although this upward bias is present in the q_m^*s, useful inferences can still be drawn. If the threshold quantile remains at 80 percent, the previously noted instances are also found to indicate concern under the two assumed distributions.[18] In addition, four new instances arise: the calibrated model in the first, second, and fifth periods, and the true model in the fifth period.

For the calibrated model in the first period depicted in Figure 2, the five observed exceptions range from about 9 percent to 27 percent more than their stated VaR estimates, which are relatively low compared to the magnitudes cited by Hendricks (1996). Thus, the "high" q_m^*s for these VaR estimates under the two assumed distributions are based more on the number of exceptions than on their magnitude. In this case, inferences based on the loss function method provide additional detail, but do not change our overall evaluation of the VaR estimates. For the calibrated model in the second period, the six exceptions are still within the yellow zone set in the MRA, but the loss function method highlights that their magnitudes, which range from 5 percent to 45 percent beyond the observed return, may be a concern.

For the fifth period, the number of exceptions are again acceptable at two, zero, and five for the true, historical, and calibrated models, respectively. Although the loss function method cannot provide additional information on the historical model due to the lack of exceptions (an acceptable outcome under this regulatory loss function), the q_m^*s for the other two models are between 80 percent and 90 percent. The reason for these high q_m^*s is that the exceptions

16. Note that, in forming the $\hat{h}_{t+1}$ series for each simulation run, an initial value $\hat{h}_1$ must be chosen. The results presented in Table 7, Panel D are based on setting $\hat{h}_1$ equal to the estimated variance of the simulated sample. An alternative specification, in which $\hat{h}_1 = \varepsilon_1^2$, generates qualitatively similar results.

17. Note that this upward bias in the q_m^*s is brought about by distributional assumptions that generate returns that, conditional on being exceptions, are not as negative as those actually observed. If the distributional assumptions were to generate returns that were generally more negative than actually observed, the bias would go in the opposite direction and cause a form of Type II error, i.e., not indicate concerns when they truly may be present. Although such distributional assumptions could be made, the general concern in practice is that observed returns are being generated from DGPs with fatter, not thinner, tails than empirically observed.

18. Note that an alternative way to conduct this type of evaluation is to recognize the upward bias imparted by the assumptions and use a higher threshold quantile, say, 90 percent. This route is complicated by the fact that the proper alternative threshold is not readily apparent. It is simpler to set the threshold quantile quite high at 80 percent and examine the flagged cases with care.

in both cases are relatively large. The true model's two exceptions are both over 50 percent of the observed returns, and the calibrated model's five exceptions range from 1 percent to 50 percent over the corresponding returns. Thus, even though both models are "acceptably accurate" under the MRA guidelines, the loss function method based on these distributional assumptions provides useful, additional (though biased) information on the performance of the VaR estimates. Regulators may use this additional information to evaluate these VaR estimates in a manner that is more directly in line with their specific concerns.

IV. CONCLUSION

As implemented in the U.S., the market risk amendment (MRA) to the Basle Capital Accord requires that large commercial banks with significant trading activities provide their regulators with VaR estimates from their own internal models. The VaR estimates are used to determine the banks' market risk capital requirements. This development clearly indicates the importance of evaluating the accuracy of VaR estimates from a regulatory perspective. In this paper, three methods for evaluating VaR estimates are discussed.

The binomial method, currently the quantitative standard in the MRA, and the interval forecast method are both based on a hypothesis-testing framework and are used to test the null hypothesis that the reported VaR estimates are "acceptably accurate," where accuracy is defined by the test conducted. As shown in the simulation exercise, the power of these tests can be low against reasonable alternative VaR models. This result does not negate their usefulness, but it does indicate that the inference drawn from them should be questioned and examined more carefully for regulatory purposes.

The loss function method is based on assigning numerical scores to the performance of the VaR estimates under a loss function that reflects the concerns of the regulators. As shown in the simulation exercise, this loss function method can distinguish between VaR estimates from the actual and alternative VaR models. Furthermore, it allows the evaluation to be tailored to specific interests that regulators may have, such as the magnitude of the observed exceptions. Although this evaluation method introduces certain biases due to necessary distributional assumptions, the analytical results provide useful additional information on the performance of the VaR estimates. Since these three methods provide complementary information, they should all be useful in the regulatory evaluation of VaR estimates.

REFERENCES

Basle Committee on Banking Supervision. 1996. "Supervisory Framework for the Use of 'Backtesting' in Conjunction with the Internal Models Approach to Market Risk Capital Requirements." Manuscript, Bank for International Settlements.

Berkowitz, J. 1999. "Evaluating the Forecasts of Risk Models." Finance and Economic Discussion Series 99-12, Federal Reserve Board of Governors.

Christoffersen, P.F. 1998. "Evaluating Interval Forecasts." *International Economic Review* 39, pp. 841–862.

Crnkovic, C., and J. Drachman. 1996. "Quality Control." *Risk* 9, pp. 139–143.

Diebold, F.X., T.A. Gunther, and A.S. Tay. 1998. "Evaluating Density Forecasts with Applications to Financial Risk Management." *International Economic Review* 39, pp. 863–883.

Diebold, F.X., and J.A. Lopez. 1996. "Forecast Evaluation and Combination." In *Handbook of Statistics, Volume 14: Statistical Methods in Finance*, eds. G.S. Maddala and C.R. Rao, pp. 241– 268. Amsterdam: North-Holland.

Federal Register. 1996. "Risk-Based Capital Standards: Market Risk" 61, pp. 47, 357–47, 378.

Granger, C.W.J., and M.H. Pesaran. 1996. "A Decision-Theoretic Approach to Forecast Evaluation." Discussion Paper 96-23, Department of Economics, University of California, San Diego.

Hendricks, D. 1996. "Evaluation of Value-at-Risk Models Using Historical Data." Federal Reserve Bank of New York *Economic Policy Review* 2, pp. 39–69.

Hendricks, D, and B. Hirtle. 1997. "Bank Capital Requirements for Market Risk: The Internal Models Approach." Federal Reserve Bank of New York *Economic Policy Review* (December) pp. 1–12.

J.P. Morgan. 1996. *RiskMetrics Technical Document* (4th ed.). New York.

Kearns, P., and A. Pagan. 1997. "Estimating the Density Tail Index for Financial Time Series." *Review of Economics and Statistics* 79, pp. 171–175.

Kupiec, P. 1995. "Techniques for Verifying the Accuracy of Risk Measurement Models." *Journal of Derivatives* 3, pp. 73–84.

Kupiec, P, and J.M. O'Brien. 1995. "A Pre-Commitment Approach to Capital Requirements for Market Risk." FEDS Working Paper #95-36, Board of Governors of the Federal Reserve System.

Lopez, J.A. 1999. "Regulatory Evaluation of Value-at-Risk Models." *Journal of Risk* 1, pp. 37–64.

Pritsker, M. 1997. "Evaluating Value-at-Risk Methodologies: Accuracy versus Computational Time." *Journal of Financial Services Research* 12, pp. 201–42.

Silverman, B.W. 1986. *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Stahl, G. 1997. "Three Cheers." *Risk* 10, pp. 67–69.