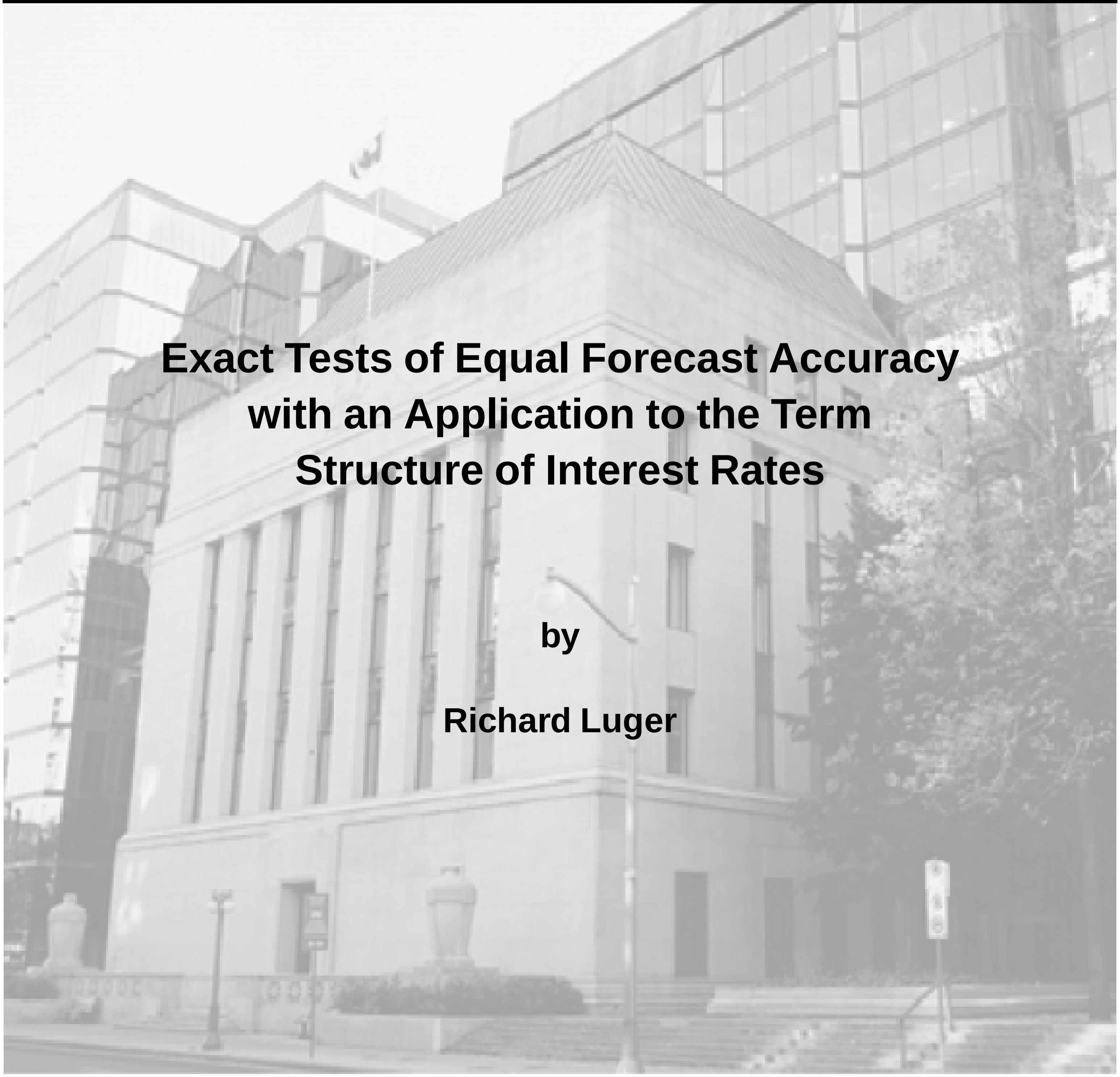
Bank of Canada Banque du Canada

Working Paper 2004-2 / Document de travail 2004-2



Exact Tests of Equal Forecast Accuracy with an Application to the Term Structure of Interest Rates

by

Richard Luger

ISSN 1192-5434

Printed in Canada on recycled paper

Bank of Canada Working Paper 2004-2

January 2004

Exact Tests of Equal Forecast Accuracy with an Application to the Term Structure of Interest Rates

by

Richard Luger

Monetary and Financial Analysis Department
Bank of Canada
Ottawa, Ontario, Canada K1A 0G9
rluger@bankofcanada.ca

The views expressed in this paper are those of the author.
No responsibility for them should be attributed to the Bank of Canada.

Contents

Abstract

The author proposes a class of exact tests of the null hypothesis of exchangeable forecast errors and, hence, of the hypothesis of no difference in the unconditional accuracy of two competing forecasts. The class includes analogues of the well-known Diebold and Mariano (1995) parametric and non-parametric test statistics. The forecast errors can be non-normal and contemporaneously correlated, and general forms of the loss function are admitted. The non-parametric distribution-free property of these new tests makes them robust to the presence of conditional heteroscedasticity, heavy tails, and outliers in the loss-differential series. These tests are used with a randomization or “Monte Carlo” resampling technique, which yields an exact and computationally inexpensive inference procedure. Simulations confirm the reliability of the new test procedure, and its power is found to be comparable with that of the size-corrected parametric Diebold-Mariano test. The test procedure is illustrated with an application to the term structure of interest rates. The application shows that exchangeable forecast errors can be found empirically even when comparing forecasts from estimated models.

JEL classification: C12, C22, C52, C53
Bank classification: Econometric and statistical methods

Résumé

L’auteur propose une classe de tests exacts de l’hypothèse nulle d’interchangeabilité des erreurs de prévision, c’est-à-dire de l’hypothèse voulant que l’exactitude inconditionnelle de deux prévisions concurrentes n’affiche aucune différence. Cette classe réunit des statistiques de test analogues à celles des tests paramétriques et non paramétriques bien connus de Diebold et Mariano (1995). Les erreurs de prévision peuvent ne pas suivre une loi normale et être corrélées de façon contemporaine. De plus, la fonction de perte peut prendre des formes générales. Le caractère non paramétrique de ces nouveaux tests, dont la statistique ne suit aucune loi prédéfinie, explique leur robustesse en présence d’hétéroscédasticité conditionnelle, de courbes de forme plus pointue et de valeurs aberrantes dans la série issue de la comparaison des fonctions de perte. Les tests s’accompagnent d’une randomisation ou d’un rééchantillonnage à la Monte-Carlo, ce qui débouche sur une procédure d’induction exacte et peu exigeante sur le plan des calculs. Les simulations confirment la fiabilité de ce nouveau test, dont la puissance se compare à celle du test paramétrique de Diebold et Mariano à niveau corrigé. Appliqué à titre illustratif à la structure des taux d’intérêt, le test montre que les erreurs de prévision peuvent s’avérer interchangeables empiriquement même lorsqu’on compare des prévisions obtenues à partir de modèles estimés.

Classification JEL : C12, C22, C52, C53
Classification de la Banque : Méthodes économétriques et statistiques

1. Introduction

An important question that occurs in time-series forecasting is how to formally compare the quality of competing forecasts. Formal comparisons attempt to assess whether differences between competing forecasts are statistically significant or simply due to sampling variability.

There are four main difficulties with formal testing: (i) forecast errors are generally not mean-zero or normally distributed, (ii) multi-step forecasts are serially correlated and heteroscedastic, (iii) competing forecasts tend to be contemporaneously correlated, and (iv) the economic loss function may be asymmetric and not correspond to the usual statistical measures, such as absolute or squared forecast error.

Let $\{(e_{1t}, e_{2t})\}_{t=1}^{T}$ be a bivariate vector time series, where the elements represent competing forecast errors. For example, each of these might be the outcome of forecasts based on judgments, surveys, smoothing, extrapolation techniques, leading indicators, time-series models, or any combination of methods. The quality of the forecasts is to be judged according to some specified loss function, $g(\cdot)$. Although the loss may depend on both the outcomes and the forecasts, it is common to assume that the loss function depends only on the forecast errors. Let $d_t = g(e_{1t}) - g(e_{2t})$ thus denote the loss differential. Typically, the null hypothesis of unconditional equal forecast accuracy is

$$H_0^{(1)} : E[d_t] = 0, \tag{1}$$

which can be interpreted as meaning that the errors associated with the two forecasts are equally costly, on average. If the null is rejected, a decision-maker would choose the forecasting method that yields the smallest loss. Given a series, $\{d_t\}_{t=1}^{T}$, of loss differentials, it is natural to base a test of (1) on the sample mean:

$$\bar{d} = \frac{1}{T}\sum_{t=1}^{T} d_t. \tag{2}$$

The Diebold and Mariano (1995) (DM) parametric test is a well-known procedure for testing the null hypothesis of no difference in the accuracy of two competing forecasts. It is given by

$$\text{DM} = \frac{\bar{d}}{\sqrt{\hat{V}(\bar{d})}}, \tag{3}$$

where $\hat{V}(\bar{d})$ is an estimate of the asymptotic variance of $\bar{d}$. Whenever an optimal forecast is produced from a proper information set, the resulting h-step forecast errors will follow a moving-average (MA) process of order $(h-1)$ of the form $e_t = \theta_0\varepsilon_t + \theta_1\varepsilon_{t-1} + ... + \theta_{h-1}\varepsilon_{t-h+1}$. Therefore, Diebold and Mariano propose estimating the variance using the truncated kernel with a bandwidth of $(h-1)$ for h-step forecasts. That estimator is computed as

$$\hat{V}(\bar{d}) = \frac{1}{T}\left[\hat{\gamma}_0 + 2\sum_{k=1}^{h-1}\hat{\gamma}_k\right], \tag{4}$$

where $\hat{\gamma}_k$ is an estimate of the kth autocovariance of d_t, given by

$$\hat{\gamma}_k = \frac{1}{T}\sum_{t=k+1}^{T}(d_t - \bar{d})(d_{t-k} - \bar{d}).$$

If the loss-differential series satisfies some regularity assumptions—such as covariance stationarity, short memory, and the existence of moments—that ensure the applicability of a central limit theorem, then the DM test statistic has an asymptotic standard normal distribution under the null hypothesis.

As the simulation experiments in Diebold and Mariano (1995) show, the normal distribution can be a very poor approximation of the DM test's finite-sample null distribution. Their results show that the DM test can have the wrong size, rejecting the null too often, depending on the degree of serial correlation among the forecast errors and the sample size, T.

Harvey, Leybourne, and Newbold (1997) (HLN) suggest that improved small-sample properties can be obtained by: (i) making a bias correction to the DM test statistic, and (ii) comparing the corrected statistic with a Student-t distribution with $(T-1)$ degrees of freedom, rather than the standard normal. The corrected statistic is obtained as

$$\text{HLN-DM} = \sqrt{\frac{T+1-2h+h(h-1)/T}{T}}\text{DM}. \tag{5}$$

Harvey, Leybourne, and Newbold advocate the use of this modified Diebold-Mariano test procedure, although the use of Student-t critical values can be justified only in the case of independent and normally distributed loss differentials.

Clark (1999) considers the size and power of several tests of equal forecast accuracy, including variants of the Diebold-Mariano test statistic with different heteroscedasticity and autocorrelation-consistent (HAC) variance estimators, such as that proposed by Newey and West (1994), which uses the Bartlett kernel and a data-determined bandwidth. His results show that all of the tests suffer size distortions in small samples, with the HLN-DM test suffering relatively the least.

This paper proposes new methods to obtain exact tests of the null hypothesis of exchangeable forecast errors and, hence, of no difference in the accuracy of two competing forecasts; i.e., if the forecast errors are exchangeable, then there is no difference in the accuracy of two competing forecasts. The proposed methods are based on the fact that exchangeable forecast errors implies symmetrically distributed loss differentials. The approach then exploits results from the theory of non-parametric statistics that show that the only tests of symmetry about zero that are valid under sufficiently general distributional assumptions, allowing for non-normality and possible heteroscedastic observations, are based on sign statistics conditional on the absolute values of the observations (see Lehmann and Stein 1949; Pratt and Gibbons 1981, 218; and Dufour 2003). In addition to the parametric DM test described above, Diebold and Mariano (1995) propose two exact (non-parametric) tests, one of which also rests on a symmetry assumption for the loss differentials. Under the null of exchangeability, the two exact tests that Diebold and Mariano (1995) propose and the tests proposed here become similar tests of symmetry about zero.

Section 2 presents the exact test procedures. It begins with a review of the two linear signed rank statistics proposed by Diebold and Mariano (1995). A class of test statistics is then introduced that includes analogues of the DM parametric and non-parametric test statistics. General forms of the loss function are admitted, and the forecast errors can be non-normal and contemporaneously correlated. In fact, the class of test statistics proposed here, while analogous to the *parametric* DM test statistic based on (2), retains the virtues of the *non-parametric* tests proposed by Diebold and Mariano: their finite-sample distributions are easily described, they are robust to departures from Gaussian conditions, and they are invariant to unknown forms of conditional heteroscedasticity.

Section 3 proposes to use these new tests with a randomization or “Monte Carlo” resampling technique that yields an exact and computationally inexpensive inference procedure. Section 4 describes the results of a small simulation study as evidence of the finite-sample performance of the proposed test procedure. Size comparisons are made with Diebold and Mariano’s (1995) parametric test and with the modified version by Harvey, Leybourne, and Newbold (1997). The power of the proposed test procedure is then compared with, and shown to be similar to, Diebold and Mariano’s size-corrected parametric test. Section 5 applies the procedure to test the predictions of the theory of the term structure of interest rates for Canada and the United States. The application shows that exchangeable forecast errors can be found empirically even when comparing forecasts from estimated models. Section 6 concludes.

2. Exact Test Procedures

In addition to the DM test in (3), Diebold and Mariano (1995) propose two exact non-parametric linear signed rank test statistics. Assuming one-step-ahead forecast errors, the first is a classical sign test given by

$$\mathrm{DM}^S = \sum_{t=1}^{T} s(d_t), \tag{6}$$

where $s(x)$ is a sign function equal to 1 when $x > 0$, and 0 otherwise. The null hypothesis under test in this case is one of median-zero loss differentials; i.e., H_0^S : median$(d_t) = 0$. That null is not quite the same as the null of no difference between median losses; i.e., median$(g(e_{1t}) - g(e_{2t})) \neq$ median$(g(e_{1t}))$ − median$(g(e_{2t}))$. Nevertheless, as Diebold and Mariano state, it has the intuitive and meaningful interpretation that $\Pr[g(e_{1t}) > g(e_{2t})] = \Pr[g(e_{1t}) < g(e_{2t})]$.

Under H_0^S, the statistic DM^S follows a binomial distribution with number of trials T and probability of success 1/2, assuming that the loss-differential series contains no zeros. The standardized version, $\mathrm{DM}_S^* = (\mathrm{DM}^S - E[\mathrm{DM}^S])/\sqrt{Var[\mathrm{DM}^S]}$, where $E[\mathrm{DM}^S] = T/2$ and $Var[\mathrm{DM}^S] = T/4$, is approximately standard normal even for relatively small values of T. If the loss differentials are symmetrically distributed about the origin, then the mean, if it exists, and the median—which always exists—are the same. In that case, (6) becomes a test of mean-zero loss differentials as written in (1).

The second non-parametric test proposed by Diebold and Mariano (1995), a Wilcoxon signed rank statistic, does indeed require the assumption of symmetrically distributed loss differentials. In the case of one-step-ahead forecast errors, the test statistic is

$$\mathrm{DM}^W = \sum_{t=1}^{T} s(d_t) R_t^+, \tag{7}$$

where R_t^+ is the rank of $|d_t|$ when $|d_1|, |d_2|, ..., |d_T|$ are placed in ascending order. Assuming that the loss differentials contain no zeros and that there are no ties among their absolute values, the statistic DM^W is distributed like the Wilcoxon variate, the distribution of which has been tabulated for various values of T; see Table A.4 in Hollander and Wolfe (1973) for $T \leq 15$. For larger values, the standard normal distribution provides a very good approximation of the standardized version $\mathrm{DM}_W^* = (\mathrm{DM}^W - E[\mathrm{DM}^W]/\sqrt{Var[\mathrm{DM}^W]}$, where $E[\mathrm{DM}^W] = T(T+1)/4$ and $Var[\mathrm{DM}^W] = T(T+1)(2T+1)/24$.

It is easy to see that one-step-ahead forecasts, if optimal, will be white noise (although not necessarily Gaussian, and hence not independent). This reflects the basic property of a white-noise series that earlier terms contain no information about later terms. If the one-step-ahead errors are not white noise, then they will at least be partially forecastable from past errors. Therefore, the original forecasts could be improved by adding to them these forecasts of the errors; see Granger and Newbold (1977, 119) and Diebold and Lopez (1996), among others, for more on the properties of optimal forecasts.

When the forecast horizon is more than just one period, it is well known that optimal h-steps-ahead forecast errors follow an MA process of order $(h-1)$. As Diebold and Mariano suggest, such forms of serial correlation can be handled via Bonferroni bounds (see also Campbell and Ghysels 1995). Assuming the loss differentials to be at most $(h-1)$-dependent, each of the following h vectors of loss differentials will be serially uncorrelated: $(d_1, d_{1+h}, d_{1+2h}, ...)$, $(d_2, d_{2+h}, d_{2+2h}, ...)$, ..., $(d_h, d_{2h}, d_{3h}, ...)$. Therefore, if h tests are performed, each with individual nominal level α/h, then Bonferroni's inequality ensures that the induced test—which consists of rejecting the null if any of the individual tests reject—has an overall level no larger than α.

The procedures proposed here test whether the forecast errors are *exchangeable*. The null hypothesis of exchangeability is formally stated as:

$$H_0^{(2)} : (e_{1t}, e_{2t}) \overset{d}{=} (e_{2t}, e_{1t}), \tag{8}$$

where $\overset{d}{=}$ stands for the equality in distribution. Forecast errors are thus deemed "equivalent" under $H_0^{(2)}$ whenever

$$\Pr[e_{1t} \leq x_1, e_{2t} \leq x_2] = \Pr[e_{2t} \leq x_1, e_{1t} \leq x_2],$$

such that the value of their joint cumulative distribution function is not affected by permutations of its arguments. It will be seen shortly that exchangeable forecast errors implies symetrically distributed loss differentials. This fact might explain why, as Diebold and Mariano note, loss differentials often appear to be symmetrically distributed in practice. Exchangeability thus provides a unifying framework in which (6) and (7) become tests of the symmetry hypothesis. Moreover, exchangeability of the forecast errors allows the construction of exact DM-type tests based on (2) that retain the robustness virtues of the sign test and the Wilcoxon signed rank test proposed by Diebold and Mariano.

The next result, given in Randles and Wolfe (1979) as Theorem 1.3.7, is useful to undersand the applications of equal-in-distribution arguments that follow.

Theorem 1. *If* $\mathbf{X} \overset{d}{=} \mathbf{Y}$ *and* $U(\cdot)$ *is a measurable function defined on the common support of* $\mathbf{X}$ *and* $\mathbf{Y}$*, then* $U(\mathbf{X}) \overset{d}{=} U(\mathbf{Y})$.

Consider a pair $(\varepsilon_{1t}, \varepsilon_{2t})$ of random variables that satisfy the exchangeability condition:

$$(\varepsilon_{1t}, \varepsilon_{2t}) \overset{d}{=} (\varepsilon_{2t}, \varepsilon_{1t}). \tag{9}$$

If ε_{1t} and ε_{2t} are independent and identically distributed (i.i.d.) random variables, then they are clearly exchangeable—but the converse is not necessarily true. For example, if ε_{1t} and ε_{2t} are jointly normal random variables, then they are exchangeable even if the correlation between them is large (Rao 1973, 196); see Galambos (1982), McCabe (1989), and Draper et al. (1993) for more on exchangeability.

Let $\theta(L) = \sum_{i=0}^{h-1} \theta_i L^i$, where L is the lag operator. Applying $\theta(L)$ and then $g(\cdot)$ to the elements on both sides of (9) implies, according to Theorem 1, that

$$(g(e_{1t}), g(e_{2t})) \overset{d}{=} (g(e_{2t}), g(e_{1t})), \tag{10}$$

where $e_{it} = \theta(L)\varepsilon_{it}$, $i = 1, 2$, represent optimal h-steps-ahead forecast errors; i.e., moving averages of exchangeable random variables are themselves exchangeable. Using $Q(X_1, X_2) = X_1$ on both sides of (10) further implies that $g(e_{1t}) \overset{d}{=} g(e_{2t})$. In that case, $E[g(e_{1t})] = E[g(e_{2t})]$, since random variables with the same distribution also have the same moments. Therefore, the truth of $H_0^{(2)}$ implies that of $H_0^{(1)}$.

Exchangeability of the forecast errors implies that the resulting loss differentials are symmetrically distributed. To see this, first note that a random variable X has a distribution that is symmetric if and only if $X \overset{d}{=} -X$. Defining $\Delta(X_1, X_2) = X_1 - X_2$ and proceeding as above, it is seen that

$$\Delta(g(e_{1t}), g(e_{2t})) \overset{d}{=} \Delta(g(e_{2t}), g(e_{1t})),$$

or

$$(g(e_{1t}) - g(e_{2t})) \overset{d}{=} -(g(e_{1t}) - g(e_{2t})),$$

so that, under $H_0^{(2)}$, the loss differentials $d_t = g(e_{1t}) - g(e_{2t})$ have symmetric distributions even if the forecast errors are contemporaneously correlated.

Even optimal one-step-ahead forecast errors never need be completely serially independent, however, because dependence can always enter through higher moments—as for example with the conditional-variance dependence of generalized autoregressive conditional heteroscedasticity (GARCH) or stochastic volatility processes—without giving rise to any conditional-mean predictability. The null hypothesis allows for such forms of dependence in higher even-numbered moments of the loss differentials. To see this, suppose that the loss differentials associated with one-step-ahead forecasts are governed by $d_t = \sigma_t \eta_t$, where $\{\eta_t\}$ is an i.i.d. sequence of random variables drawn from a symmetric distribution (such as a standard normal or Student-t distribution). Let $I_t = (d_t, d_{t-1}, ...)$, and suppose that, conditional on I_{t-1}, σ_t and η_t are independent. Consider the conditional distribution of η_T given η_1^{T-1}, where $\eta_1^t = (\eta_1, ..., \eta_t)$. Since η_t are i.i.d. and symmetric, $(\eta_T | \eta_1^{T-1}) \overset{d}{=} (-\eta_T | \eta_1^{T-1})$. It follows that $(\sigma_T \eta_T | \eta_1^{T-1}) \overset{d}{=} (-\sigma_T \eta_T | \eta_1^{T-1})$, which in turn implies that $(\sigma_T \eta_T, \eta_1^{T-1}) \overset{d}{=} (-\sigma_T \eta_T, \eta_1^{T-1})$. This argument can be repeated recursively to find that the unconditional distribution of the loss differentials is indeed multivariate symmetric. The conditional variance need not be finite or even follow a stationary process. In fact, no restrictions are placed on the degree of heterogeneity and dependence of even-numbered moments about the origin.

The test procedures proposed here are constructed on the basis of the multivariate symmetry of the loss differentials that results under $H_0^{(2)}$. Consider the general case of optimal h-steps-ahead forecast errors following MA processes of order $(h-1)$ that satisfy (10) with associated loss differentials $d_t = g(e_{1t}) - g(e_{2t})$ for $t = 1, ..., T$. Suppose that T/h is an integer and consider the h vectors each containing T/h elements,

$$\begin{aligned} D_1 &= (d_1, d_{1+h}, d_{1+2h}, d_{1+3h}, ..., d_{T-h+1}), \\ D_2 &= (d_2, d_{2+h}, d_{2+2h}, d_{2+3h}, ..., d_{T-h+2}), \\ &\vdots \\ D_h &= (d_h, d_{2h}, d_{3h}, d_{4h}, ..., d_T), \end{aligned} \tag{11}$$

where the elements of D_i, $i = 1, ..., h$, are separated by $(h-1)$ periods. Let T_i denote the collection of indices defining D_i. The multivariate symmetry that results under $H_0^{(2)}$ implies that

$$(d_i, d_{i+h}, ..., d_{T-h+i}) \overset{d}{=} (-d_i, d_{i+h}, ..., d_{T-h+i}) \overset{d}{=} \cdots \overset{d}{=} (-d_i, -d_{i+h}, ..., -d_{T-h+i}), \tag{12}$$

where all $2^{T/h}$ such terms appear in this string of equalities in distribution. If covariances are finite, then multivariate symmetry implies that $E[d_t d_s] = 0$, for any $d_t, d_s \in D_i$, $t \neq s$

(Randles and Wolfe 1979, Lemma 1.3.28). The existence of any moments, however, need not be assumed for the validity of the proposed test procedure.

Define the sign function:

$$\tilde{s}(x) = \begin{cases} 1 & \text{if } x > 0, \\ -1 & \text{if } x < 0, \\ \pm 1 & \text{if } x = 0, \end{cases} \tag{13}$$

where ±1 means that the sign is chosen randomly with probability 1/2. When $x = 0$, a simple way to implement the sign function is to generate a random variate, say z, from any continuous symmetric distribution—a standard normal, for example—and then assign the ratio $z/|z|$ as the sign of x. With this definition of the sign function, exact test procedures can be obtained even when the loss-differential series contains zeros; see Pratt and Gibbons (1981, 160).

Consider then the class of statistics defined by

$$SF(D_i) = \sum_{t \in T_i} \tilde{s}(d_t) f(|d_t|), \tag{14}$$

where $f(\cdot)$ is a non-negative non-decreasing function of the absolute values of the loss differentials. The next result is the basic building block for constructing hypothesis tests that have a non-parametric distribution-free property. This property ensures robustness against the presence of conditional heteroscedasticity, heavy tails, and outliers in the loss-differential series.

Theorem 2. *Under $H_0^{(2)}$, any statistic defined by (14) has the property, conditional on $|D_i| = (|d_i|, |d_{i+h}|, ..., |d_{T-h+i}|)$, that*

$$\sum_{t \in T_i} \tilde{s}(d_t) f(|d_t|) \overset{d}{=} \sum_{t \in T_i} S_t f(|d_t|),$$

where $S_i, S_{i+h}, ..., S_{T-h+i}$ are mutually independent random variables, such that $\Pr[S_t = 1] = \Pr[S_t = -1] = 1/2$ for $t \in T_i$.

Proof. Given the realizations $|D_i|$, multivariate symmetry implies that the $2^{T/h}$ possible T/h vectors

$$(\pm|d_i|, \pm|d_{i+h}|, ..., \pm|d_{T-h+i}|)$$

are equally likely, where $\pm|d_t|$ means that $|d_t|$ is independently assigned either a positive or negative sign with equal probability. Therefore, conditional on $|D_i|$, the $2^{T/h}$ possible values of $SF(D_i)$ derived from the $2^{T/h}$ possible sign assignments represented by

$$SF(\pm|D_i|)$$

are also equally likely. □

Based on Theorem 1, an α/h-level conditional test can be performed as follows. If large absolute values of the test statistic $SF(D_i)$ are more probable under the alternative, the null hypothesis is rejected when the observed value of $|SF(D_i)|$ falls in a set, $C_i(\alpha/h)$, that contains the $2^{T/h}\alpha/h$ largest absolute values of the test statistic $SF(D_i)$ that can be obtained from the class of all sign assignments.

Critical regions constructed from the equally likely property established in Theorem 1 are conditional ones; i.e., they depend on $|D_i|$, which is obtained once the data have been observed. Consider the conditional probability of a Type I error for such a test, which may be written as

$$E\left[\mathbb{I}(SF(D_i) \in C_i(\alpha/h)) \ \mid \ |D_i|\right] \leq \alpha/h, \tag{15}$$

where $\mathbb{I}(\cdot)$ is the indicator function. By taking expectations on both sides of (15), such a test is seen to also have level α/h unconditionally.

Consider the Bonferroni decision rule that consists of rejecting $H_0^{(2)}$ when it has been rejected by at least one of the tests based on an $SF(D_i)$ statistic. The critical region corresponding to this decision rule is $\bigcup_{i=1}^{h} C_i(\alpha/h)$ and therefore, under $H_0^{(2)}$,

$$\Pr\left[\bigcup_{i=1}^{h} SF(D_i) \in C_i(\alpha/h)\right] \leq \sum_{i=1}^{h} \Pr\left[SF(D_i) \in C_i(\alpha/h)\right] \leq \alpha,$$

so that the induced test has an overall significance level no larger than α.

Determination of the sets $C_i(\alpha/h)$ by direct counting would be impractical in most cases. Section 3 illustrates how a Monte Carlo resampling technique can be used to perform exact inference without the need to enumerate the entire randomization distribution. Section 3 also describes how the sample-split approach can be used in conjunction with the resampling technique to obtain an exact Monte Carlo test without relying on a Bonferroni rule. The Monte Carlo test is particularly useful whenever h is large and α/h is deemed to be too small relative to the desired overall significance level.

The class of statistics defined by (14) includes classical linear signed rank statistics of the form

$$SR(D_i) = \sum_{t \in T_i} 0.5(\tilde{s}(d_t) + 1) a_{T/h}(\tilde{R}_t^+), \tag{16}$$

where the set of scores $a_{T/h}(t)$, $t = 1, ..., T/h$, satisfy $0 \leq a_{T/h}(1) \leq ... \leq a_{T/h}(T/h)$ with $a_{T/h}(T/h) > 0$, and $\tilde{R}_t^+$ is the rank of $|d_t|$ when the pairs $(|d_i|, u_i), (|d_{i+h}|, u_{i+h}), ..., (|d_{T-h+i}|, u_{T-h+i})$—with u_i being i.i.d. draws from a continuous uniform distribution—are arranged in lexicographic order:

$$(|d_i|, u_i) < (|d_j|, u_j) \Leftrightarrow \{|d_i| < |d_j| \text{ or } (|d_i| = |d_j| \text{ and } u_i < u_j)\}.$$

The lexicographic ordering ensures that the ranks used in (16) are well defined when two or more loss differentials have the same absolute value.

Note that, conditional on $(|d_i|, u_i), (|d_{i+h}|, u_{i+h}), ..., (|d_{T-h+i}|, u_{T-h+i})$, the vector of scores used in (16) is a fixed permutation of $(a_{T/h}(1), a_{T/h}(2), ..., a_{T/h}(T/h))$. Therefore, the distribution of any statistic of the form (16), derived under the equally likely principle in Theorem 1, also holds unconditionally, since it does not depend on $(|d_i|, u_i), (|d_{i+h}|, u_{i+h}), ..., (|d_{T-h+i}|, u_{T-h+i})$.

Corollary 2.1 *Under $H_0^{(2)}$, any statistic defined by (16) has the property that*

$$\sum_{t \in T_i} 0.5(\tilde{s}(d_t) + 1) a_T(\tilde{R}_t^+) \overset{d}{=} \sum_{t=1}^{T/h} B_t a_{T/h}(t),$$

where $B_1, B_2, ..., B_{T/h}$ are mutually independent uniform Bernoulli variables, such that $\Pr[B_t = 1] = \Pr[B_t = 0] = 1/2$ *for* $t = 1, 2, ..., T/h$.

Within the class of statistics defined by (16), consider the sign statistic, which is obtained from the constant score function $a_{T/h}(t) = 1$:

$$S(D_i) = \sum_{t \in T_i} 0.5(\tilde{s}(d_t) + 1), \tag{17}$$

and the Wilcoxon signed rank statistic

$$W(D_i) = \sum_{t \in T_i} 0.5(\tilde{s}(d_t) + 1) \tilde{R}_t^+, \tag{18}$$

obtained with $a_{T/h}(t) = t$. Diebold and Mariano (1995) also propose these two nonparametric test statistics assuming i.i.d. loss differentials. The following result establishes that the statistics defined in (17) and (18) have the usual distributions under far more general distributional assumptions.

Corollary 2.2 *Under* $H_0^{(2)}$*:*
(i) The statistic $S(D_i)$*, defined by (17), is distributed according to* $B(T/h, 1/2)$*, a binomial distribution with number of trials* T/h *and probability of success* $1/2$.
(ii) The statistic $W(D_i)$*, defined by (18), is distributed like* $\sum_{t=1}^{T/h} tB_t$.

As stated above, the distribution of the Wilcoxon variate, $W(D_i)$, has been tabulated for various sample sizes and, following standard results described by Randles and Wolfe (1979, Section 10.2), it can be shown that the standardized linear signed rank statistic,

$$\left[SR(D_i) - \frac{1}{2}\sum_{t=1}^{T/h} a_{T/h}(t) \right] \Bigg/ \sqrt{\frac{1}{4}\sum_{t=1}^{T/h} a_{T/h}^2(t)}\ ,$$

has a limiting standard normal distribution.

A clear advantage of the sign and Wilcoxon test statistics is that they can be used by simply referring to standard statistical tables to find appropriate critical values, thereby avoiding the need for simulations. These non-parametric statistics based on signed ranks, however, are expected to be less powerful than other members of the class defined in (14), such as $\sum_{d_t \in D_i} \tilde{s}(d_t)|d_t|$, which exploits all the information contained in the signs and the absolute values of the loss differentials. Section 3 describes how to conduct inference based on such a test statistic.

3. Monte Carlo Test Procedure

In only a few cases do test statistics defined by (14) have perfectly tabulated null distributions. Examples include the non-parametric linear signed rank test statistics in Corollary 2. This section describes how to find the null distribution of any statistic defined by (14) based on its characterization in Theorem 1.

Generation of the entire randomization distribution of a test statistic defined by (14), by a complete enumeration of all possible sign assignments, is computationally prohibitive for sample sizes typical in applied work. The computational burden of finding tail probabilities can be reduced by drawing samples from the sign-randomization distribution and computing the value of (14) each time. The relative frequencies of these values comprise the simulated sign-randomization distribution. The Monte Carlo procedure of Dwass (1957) provides a simple method to obtain the desired significance level and a precise p-value, without performing a large number of draws. The construction of such a Monte Carlo test is illustrated with the absolute value of

$$\text{MC-DM} = \sum_{t=1}^{T} \tilde{s}(d_t)|d_t|, \tag{19}$$

as a two-sided test of the null hypothesis of exchangeable one-step-ahead forecast errors.

Let $|\text{MC-DM}_B|$ denote the absolute value of MC-DM computed from the original sample $(d_1, d_2, ..., d_T)$, and let $|\text{MC-DM}_b|$, $b = 1, ..., B-1$, denote those obtained by randomly sampling the sign-randomization distribution; i.e., $|\text{MC-DM}_b|$ is the absolute value of MC-DM computed from $(\pm|d_1|, \pm|d_2|, ..., \pm|d_T|)$. Note that the sign-randomization distribution is discrete, so that ties among the randomly sampled statistics have a non-zero probability of occurrence. To break ties, draw B variates U_i, $i = 1, ..., B$, from a continuous uniform distribution independently of the MC-DM_b's and arrange the pairs $(|\text{MC-DM}_1|, U_1), (|\text{MC-DM}_2|, U_2), ..., (|\text{MC-DM}_B|, U_B)$ according to lexicographic order:

$$\begin{aligned} &(|\text{MC-DM}_i|, U_i) < (|\text{MC-DM}_j|, U_j) \Leftrightarrow \\ &\qquad [|\text{MC-DM}_i| < |\text{MC-DM}_j| \text{ or } (|\text{MC-DM}_i| = |\text{MC-DM}_j| \text{ and } U_i < U_j)]\,. \end{aligned}$$

Let $\tilde{R}_B$ denote the rank of $(|\text{MC-DM}_B|, U_B)$ in the lexicographic ordering, which is easily computed as:

$$\tilde{R}_B = 1 + \sum_{i=1}^{B-1} \mathbb{I}\left(|\text{MC-DM}_B| > |\text{MC-DM}_i|\right) + \sum_{i=1}^{B-1} \mathbb{I}\left(|\text{MC-DM}_B| = |\text{MC-DM}_i|\right) \times \mathbb{I}\left(U_B > U_i\right),$$

where $\mathbb{I}(\cdot)$ is again the indicator function. If αB is an integer, then

$$\Pr\left[\tilde{R}_B \geq B - \alpha B + 1\right] = \alpha,$$

such that $p_B = (B - \tilde{R}_B + 1)/B$ can be interpreted as a randomized p-value, which can be used to perform a test with size α.

For a given, possibly small, number of random draws, the Monte Carlo procedure allows the size of the test to be controlled with exactness. This feature stands in sharp contrast with bootstrap test procedures, which generally are valid only asymptotically. As B increases without bound, the inference based on the Monte Carlo procedure becomes equivalent to that based on the equivalent non-randomized procedure; see Dufour (2000) and Dufour and Khalaf (2001).

Consider now the case of h-steps-ahead forecasts governed by an MA$(h-1)$ process. Once the associated sample of loss differentials has been split into subsamples according to (11), the Monte Carlo technique can be used to obtain combined inference across subsamples without relying on a Bonferroni rule. To that end, consider the following statistic:

$$\text{MC-DM}^{\max} = \max_{1 \le i \le h} |\text{MC-DM}(D_i)|,$$

where MC-DM(D_i) is the value of MC-DM computed on the subsample D_i, $i = 1, ..., h$. Given that the statistics MC-DM(D_i) are *jointly* pivotal, this criterion is statistically equivalent to choosing the statistic with the smallest two-sided p-value. This type of criterion is derived from the logical equivalence that the null of equal forecast accuracy is true if and only if it holds true over each subsample; the null will be rejected if at least one of the individual tests is significant. The steps of the combined test procedure are:

(i) compute the value of MC-DM$_B^{\max}$ based on the original h subsamples $D_1, D_2, ..., D_h$;

(ii) compute MC-DM$_b^{\max}$ $= \max_{1 \le i \le h} |\text{MC-DM}(\pm D_i)|$, $b = 1, ..., B-1$, where $\pm D_i = \{\pm d_i, \pm d_{i+h}, \pm d_{i+2h}, ..., \pm d_{T-h+i}\}$.

The Monte Carlo p-value of the combined test is given by $p_B = (B - \tilde{R}_B + 1)/B$, where $\tilde{R}_B$ now denotes the rank of (MC-DM$_B^{\max}, U_B$) once the pairs (MC-DM$_1^{\max}, U_1$), (MC-DM$_2^{\max}, U_2$), ..., (MC-DM$_B^{\max}, U_B$) are placed in ascending lexicographic order.

4. Simulation Experiment

The simulation experiment in Diebold and Mariano (1995) reveals that the DM test can be seriously oversized, especially when the loss differentials are serially correlated and when the sample size is small. Harvey, Leybourne, and Newbold (1997) and Clark (1999) provide further evidence of this phenomenon. This section reports the results of a small-scale simulation experiment to compare the relative performance of the test proposed here with the Diebold and Mariano (1995) original test and with its variant HLN-DM proposed by Harvey, Leybourne, and Newbold (1997).

The experiment design described herein follows closely the design described in the studies identified above. The experiment consists of drawing realizations of the bivariate forecast-error process $\{(e_{1t}, e_{2t})\}_{t=1}^{T}$, with varying degrees of contemporaneous and serial correlation. The forecast errors are generated by

$$\begin{pmatrix} e_{1,t} \\ e_{2,t} \end{pmatrix} = \begin{pmatrix} \frac{1+\theta L}{\sqrt{1+\theta^2}} & 0 \\ 0 & \frac{1+\theta L}{\sqrt{1+\theta^2}} \end{pmatrix} \begin{pmatrix} \sqrt{k} & 0 \\ \rho & \sqrt{1-\rho^2} \end{pmatrix} \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix},$$

where L is the lag operator, and $\varepsilon_{i,t}$, $i = 1, 2$, $t = 0, 1, ..., T$, are $IIN(0, 1)$. The null hypothesis is represented by $k = 1$, in which case the forecast errors e_{1t}, e_{2t} have equal variances. Scaling by $\sqrt{1+\theta^2}$ makes the unconditional variance equal to one under the null. The MA parameter, θ, takes the values $0, 0.5$, and 0.9. With this design, setting $\theta = 0$ yields one-step-ahead forecast errors, while two-steps-ahead forecast errors are obtained whenever $\theta \neq 0$. Values of the contemporaneous correlation coefficient $\rho = 0, 0.5, 0.9$, and sample sizes $T = 8, 16, 32, 64$ are considered. As in previous studies, a quadratic loss function is used. The results reported in Tables 1 and 2 are based on 5,000 replications of each data-generating configuration; the Monte Carlo tests are implemented with $B = 100$.

Following Clark (1999), data generated using $\theta = 0$ are treated as one-step-ahead forecasts, and in that case the DM test statistic is evaluated without any lags in its variance estimate. In their original simulation study, Diebold and Mariano (1995) treat this case as two-steps-ahead forecasts and use one lag when computing the variance estimate. When $\theta \neq 0$, the data are treated as two-steps-ahead forecasts, as in Diebold and Mariano; the MC-DM$^{\max}$ test procedure is applied in those cases.

Table 1 evaluates size based on a nominal level of 5 per cent. It is clear from the results in Table 1 that both the DM test and its variant HLN-DM suffer size distortions in small samples. For a given sample size, those size distortions are more severe when the forecast horizon increases and when the forecast errors are contemporaneously correlated. In some of those cases, the DM test has rejection rates of over 20 per cent, more than four times the nominal level. As Clark (1999) notes, the tendency of the DM test to overreject as the forecast horizon increases likely results from the difficulty in obtaining a precise estimate of the variance used in the test statistic. In general, greater serial correlation tends to make that variance estimate more imprecise.

The best performance of the DM and HLN-DM tests occurs in the case of one-step-ahead serially uncorrelated forecasts. In that case, as expected, the HLN-DM test has empirical levels that correspond closely to the stated level, although for the smallest sample size, $T = 8$, it does appear somewhat conservative. As the sample size increases, the empirical levels of both DM and HLN-DM tend to the nominal level.

On the other hand, the Monte Carlo Diebold-Mariano test has the stated level. Its performance is invariant to the sample size, T, and combinations of θ and ρ, since each of these features is implicitly accounted for by the Monte Carlo test procedure.

Table 2 compares the empirical power of the DM and Monte Carlo DM test procedures when $k = 0.75, 0.5$. The power results for the DM test are based on size-corrected critical values. For each parameter configuration, size-corrected power is calculated as the frequency with which the DM test statistic exceeds the empirical critical values in the corresponding size experiment. Since the HLN-DM test statistic is simply the DM test statistic multiplied by a constant adjustment factor, these two tests have exactly the same size-corrected power, and therefore the HLN-DM test is not considered in Table 2.

The overall picture that emerges from Table 2 is one of good power with small differences between the size-corrected DM and Monte Carlo DM tests. As expected, power increases with both ρ and T.

In the case of one-step-ahead forecasts ($\theta = 0$), the DM and MC-DM tests have virtually the same power. In the case of two-steps-ahead forecasts ($\theta \neq 0$), the MC-DM$^{\max}$ test procedure appears slightly less powerful. This is not surprising, given that some information is lost when the sample is split in two. It is important, however, to emphasize that the size-corrected DM test is not a feasible test in practice. It is merely used here as a theoretical benchmark to which the Monte Carlo tests may be compared.

5. Empirical Illustration

The tests of equal forecast accuracy are illustrated with an application to Canadian and U.S. three-month and six-month treasury bills. According to the theory of the term structure of interest rates, a longer-term interest rate can be analyzed as a weighted sum of current and expected future short rates and a constant risk premium. Let $r_t^{(3)}$ be the three-month rate and $r_t^{(6)}$ be the six-month rate. According to the strict form of the theory,

$$r_t^{(6)} = \theta + 0.5 r_t^{(3)} + 0.5 E_t r_{t+1}^{(3)},$$

where $E_t r_{t+1}^{(3)}$ denotes the time-t market forecast of $r_{t+1}^{(3)}$. If expectations are formed efficiently, then the overlapping forecast errors,

$$r_{t+1}^{(3)} - E_t r_{t+1}^{(3)} = r_{t+1}^{(3)} - 2 r_t^{(6)} + r_t^{(3)} + 2\theta,$$

should be independent of all information available to the market at time t. Furthermore, these errors should be serially uncorrelated at lags greater than two when observed at the monthly frequency.

Mankiw and Summers (1984) test the expectations theory at the short end of the term structure with strikingly negative results. Their regression results, along with the more general ones of Campbell and Shiller (1991), indicate that future rates move in the opposite direction from that predicted by the theory.

Bekaert, Hodrick, and Marshall (1997) demonstrate that regression-based in-sample tests of the expectations hypothesis are severely biased in finite samples. In particular, they show that the high persistence and heteroscedasticity of short-term interest rates induce extreme bias and dispersion into the finite-sample distributions of test statistics. They conclude that the inference based on the asymptotic distribution of these test statistics is unreliable.

Campbell and Dufour (1997) propose several variants of signed rank test statistics to test orthogonality conditions. Those tests have known finite-sample distributions and they allow for non-normal and possibly heteroscedastic observations. In contrast to the usual literature, Campbell and Dufour find for Canadian data that the expectations theory cannot be rejected once their more correct non-parametric inference procedures are used. Such studies typically test the hypothesis that $\beta_1 = 2$ and $\beta_2 = -1$ in regressions of the form

$$r^{(3)}_{t+1} = \beta_0 + \beta_1 r^{(6)}_t + \beta_2 r^{(3)}_t + \varepsilon_{t+1},$$

or in reparameterized forms that test the exclusion of the spread $(r^{(6)}_t - r^{(3)}_t)$ as an explanatory right-hand-side variable.

In general, traditional methods assess a predictive model based on its ability to “fit” the same observations used to estimate the model. Those approaches can be unreliable when the underlying data-generating process has changed over the observation period. Here, the goal is to perform out-of-sample tests that can mitigate the effects of data heterogeneity through rolling-window estimation of model parameters. The out-of-sample predictions of the theory can be tested by comparing the constrained forecast errors,

$$e_{1t+1}(\hat{\beta}_{0t}) = r^{(3)}_{t+1} - \tilde{\beta}_{0t} - (2r^{(6)}_t - r^{(3)}_t), \tag{20}$$

against the unconstrained ones,

$$e_{2t+1}(\hat{\beta}_{0t}, \hat{\beta}_{1t}, \hat{\beta}_{2t}) = r^{(3)}_{t+1} - \hat{\beta}_{0t} - (\hat{\beta}_{1t} r^{(6)}_t - \hat{\beta}_{2t} r^{(3)}_t), \tag{21}$$

where the dependence on time-t parameter estimates is made explicit. Note that (20) allows for a possible time-varying risk premium, while all the parameters appearing in

(21) can potentially vary over time. The null hypothesis of exchangeable forecast errors in this case becomes

$$H_0^{(2)} : (e_{1t}(\tilde{\beta}_{0t-1}), e_{2t}(\hat{\beta}_{0t-1})) \overset{d}{=} (e_{2t}(\hat{\beta}_{0t-1}), e_{1t}(\tilde{\beta}_{0t-1})), \tag{22}$$

when expressed in terms of time-$(t-1)$ parameter estimates. If $H_0^{(2)}$ is true, then so is

$$H_0^{(1)} : E\left[g(e_{1t}(\tilde{\beta}_{0t-1})) - g(e_{2t}(\hat{\beta}_{0t-1}, \hat{\beta}_{1t-1}, \hat{\beta}_{2t-1}))\right] = 0.$$

Intuitively, equal forecast accuracy as stated in $H_0^{(1)}$ asks whether the two forecasting procedures, which include the choice of estimation method and estimation window, produce equally accurate forecasts, on average, over a finite period of time. Giacomini and White (2003) construct asymptotic tests for the more restrictive hypothesis of conditional predictive ability, which asks the same question conditionally on time-$(t-1)$ information. The DM test can be seen as a particular case of their framework when no conditioning information is used.

The predictive model used in (20) is nested within the one used in obtaining (21). With nested models, it is possible that the limiting distribution of the Diebold-Mariano test statistic could differ from normality because, under the correctly specified null, the forecast errors are asymptotically identical and therefore perfectly correlated. McCracken (1999) and Clark and McCracken (2001) derive the asymptotic (context-specific) distributions of several tests of equal forecast accuracy and encompassing for nested models under several maintained assumptions. Of crucial importance is the assumption that the data conform to the restrictions of the nested model. If the alternative is the true model, or if both models are false, then the forecast errors will not necessarily tend to be perfectly correlated as the sample size grows without bound. Moreover, when the size of the estimation sample remains finite as the size of the prediction sample grows, parameter estimates are prevented from reaching their probability limits and the Diebold-Mariano test remains asymptotically valid even for nested models, under some regularity assumptions (Giacomini and White 2003). Essentially, this means that model parameters are estimated using a rolling window of data, rather than an expanding one. On the other hand, the procedures proposed here will always yield a valid finite-sample inference regardless of the choice of estimation method and estimation window.

The data, presented in the appendix, are monthly treasury bill secondary market rates for the period covering March 1993 to March 2003, for a total of 121 observations. Model parameters are estimated by least squares using a rolling window of length 12, 24, 36, 48, and 60 months, resulting in out-of-sample loss differentials series of length 108, 96, 84, 72, and 60, respectively, once the first lag is allowed for.

As alluded to above, serial correlation is induced by overlapping expectations. The out-of-sample loss differential series are therefore divided into three subsamples taken at three-month intervals to apply the tests of equal forecast accuracy with the Bonferroni rule. The subsample errors are treated as one-step-ahead forecast errors when evaluating the Diebold-Mariano test and the Harvey-Leybourne-Newbold variant. In the case of the full sample, the errors are three-steps-ahead forecast errors, and those tests are computed using the truncated kernel with a bandwidth of two. The Monte Carlo tests, MC-DM for the subsamples and MC-DM$^{\max}$ for the full sample, are implemented following the procedure described in section 3 with $B = 2000$.

The results are reported in Tables 3 and 4, where the entries are two-sided p-values in percentages of the null hypothesis of equal forecast accuracy between the models in (20) and (21) for Canada and the United States, respectively. Results are reported for both mean squared errors (MSE) and mean absolute errors (MAE) as mean loss criteria. Rejections at the conventional 5 per cent level are indicated by an asterisk; the decision rule for the bounds tests is to reject the null if any of the individual subsample tests rejects at the 5/3 per cent level.

The most striking result is the contrast between Canada and the United States. While the out-of-sample predictions of the expectations theory are rejected in some instances in the case of Canada—depending on which test is considered, and especially the length of the estimation window—they are never rejected in the case of the United States.

Table 3 shows that the expectations theory finds more support the shorter the estimation window. The failure of the theory for longer estimation windows is indicative of a non-constant underlying process generating Canadian interest rates over the sample period; see Clements and Hendry (1998) for a related discussion on forecast failure resulting from structural breaks. In general, the asymptotic tests concur with the exact Monte Carlo tests, which is a non-rejection of $H_0^{(2)}$ as stated in (22). This clearly illustrates that exchangeable forecast errors can be found empirically even when forecasts from estimated models are compared.

6. Conclusion

The test procedure developed in this paper tests the null hypothesis of exchangeable forecast errors and, hence, offers a solution to the potential over-rejection problem associated with standard tests of equal forecast accuracy, such as the parametric Diebold and Mariano (1995) test.

These exact non-parametric distribution-free tests are based on the independence, under the null hypothesis, of the absolute value of a loss differential and its sign meaning that, given $|d_t|$, the two observation values $+|d_t|$ and $-|d_t|$ are equally likely. Therefore, the tests introduced are conditional ones, created after the loss differentials $d_1, d_2, ..., d_T$ have been observed. The associated test procedure has an overall level of α, however, because the critical region is constructed so that the conditional probability of rejecting the null, when it is really true, is α. When the absolute values $|d_1|, |d_2|, ..., |d_T|$ are replaced with some non-negative scores, $a_T(1), a_T(2), ..., a_T(T)$, that are ordered among themselves in the same manner as $|d_1|, |d_2|, ..., |d_T|$, the resulting tests have a critical region that can be tabulated once and for all. Hence, such a test is no longer a conditional one. The sign test and the Wilcoxon signed rank test, which are also considered by Diebold and Mariano (1995), are tests of this type.

To complement the results that established the exactness of the new tests, the results of a simulation experiment have shown that the inference procedure has respectable power relative to the parametric Diebold-Mariano test.

A clear advantage of the tests described in this paper is that they are invariant to deviations from standard assumptions such as those of normality, homoscedasticity, and the existence of moments required for the validity of many parametric test methods. The non-parametric distribution-free property of the new tests makes them robust to the presence of conditional heteroscedasticity, heavy tails, and outliers in the loss-differential series.

References

Bekaert, G., R.J. Hodrick, and D.A. Marshall. 1997. “On Biases in Tests of the Expectations Hypotesis of the Term Structure of Interest Rates.” *Journal of Financial Economics* 44: 309-48.

Campbell, B. and J-M. Dufour. 1997. “Exact Nonparametric Tests of Orthogonality and Random Walk in the Presence of a Drift Parameter.” *International Economic Review* 38: 151-73.

Campbell, B. and E. Ghysels. 1995. “Is the Outcome of the Federal Budget Process Unbiased and Efficient? A Nonparametric Assessment.” *Review of Economics and Statistics* 77: 17-31.

Campbell, J.Y. and R.J. Shiller. 1991. “Yield Spreads and Interest Rate Movements: A Bird’s Eye View.” *Review of Economic Studies* 58: 495-514.

Clark, T.E. 1999. “Finite-sample Properties of Tests of Equal Forecast Accuracy.” *Journal of Forecasting* 18: 489-504.

Clark, T.E. and M.W. McCracken. 2001. “Tests of Equal Forecast Accuracy and Encompassing for Nested Models.” *Journal of Econometrics* 105: 85-110.

Clements, M.P. and D.F. Hendry. 1998. *Forecasting Economic Time Series.* Cambridge: University Press.

Diebold, F.X. and J.A. Lopez. 1996. “Forecast Evaluation and Combination.” In *Handbook of Statistics, Vol. 14: Statistical Methods in Finance*, edited by G.S. Maddala and C.R. Rao. Amsterdam: North-Holland.

Diebold, F.X. and R.S. Mariano. 1995. “Comparing Predictive Accuracy.” *Journal of Business and Economic Statistics* 13: 253-63.

Draper, D., J.S. Hodges, C.L. Mallows, and D. Pregibon. 1993. “Exchangeability and Data Analysis.” *Journal of the Royal Statistical Society, A.* 156: 9-37.

Dufour, J-M. 2000. “Monte Carlo Tests with Nuisance Parameters: A General Approach to Finite-Sample Inference and Nonstandard Asymptotics in Econometrics.” Discussion Paper, C.R.D.E., Université de Montréal.

Dufour, J-M. 2003. "Identification, Weak Instruments, and Statistical Inference in Econometrics." *Canadian Journal of Economics* 36: 767-808.

Dufour, J-M. and L. Khalaf. 2001. "Monte Carlo Test Methods in Econometrics." In *Companion to Theoretical Econometrics,* edited by B. Baltagi. Oxford: Basil Blackwell.

Dwass, M. 1957. "Modified Permutation Tests for Nonparametric Hypotheses." *Annals of Mathematical Statistics* 28: 181-87.

Galambos, J. 1982. "Exchangeability." In *Encyclopedia of Statistical Sciences,* Vol. 2, edited by S. Kotz, N.L. Johnson, and C.B. Read. New York: Wiley.

Giacomini, R. and H. White. 2003. "Tests of Conditional Predictive Ability." Working Paper No. 572, Boston College.

Granger, C.W.J. and P. Newbold. 1977. *Forecasting Economic Time Series.* New York: Academic Press.

Harvey, D., S. Leybourne, and P. Newbold. 1997. "Testing the Equality of Prediction Mean Squared Errors." *International Journal of Forecasting* 13: 281-91.

Hollander, M. and D.A. Wolfe. 1973. *Nonparametric Statistical Methods.* New York: Wiley.

Lehmann, E.L. and C. Stein. 1949. "On the Theory of Some Nonparametric Hypotheses." *Annals of Mathematical Statistics* 20: 28-45.

Mankiw, N.G. and L. Summers. 1984. "Do Long-term Interest Rates Overreact to Short-term Interest Rates?" Brookings Papers on Economic Activity 1984: 223-42.

McCabe, B.P.M. 1989. "Misspecification Tests in Econometrics Based on Ranks." *Journal of Econometrics* 40: 261-78.

McCracken, M.W. 1999. "Asymptotics for Out-of-sample Tests of Causality." Manuscript, Louisiana State University.

Newey, W.K. and K.D. West. 1994. "Automatic Lag Selection in Covariance Matrix Estimation." *Review of Economic Studies* 61: 631-53.

Pratt, J.W. and J.D. Gibbons. 1981. *Concepts of Nonparametric Theory.* New York: Springer-Verlag.

Randles, R.H. and D.A. Wolfe. 1979. *Introduction to the Theory of Nonparametric Statistics*. New York: Wiley.

Rao, C.R. 1973. *Linear Statistical Inference and its Applications.* New York: Wiley.

Table 1. Empirical Test Size in Percentage: $k = 1$

	$\rho = 0$				$\rho = 0.5$				$\rho = 0.9$			
Test	T=8	T=16	T=32	T=64	T=8	T=16	T=32	T=64	T=8	T=16	T=32	T=64
					$\theta = 0$ (one-step-ahead forecasts)							
DM	10.08	6.94	5.88	5.80	8.76	7.04	5.36	5.00	9.64	7.02	5.88	5.30
HLN-DM	3.24	4.30	5.36	5.66	2.96	4.26	4.94	4.82	3.12	3.98	5.44	5.14
MC-DM	5.20	5.20	4.84	5.44	4.72	5.24	4.54	4.66	4.94	4.88	5.10	4.98
					$\theta = 0.5$ (two-steps-ahead forecasts)							
DM	21.54	13.26	8.50	6.64	23.70	12.38	7.86	6.82	23.50	12.34	8.68	6.22
HLN-DM	11.48	7.36	7.02	5.92	12.14	6.92	6.70	6.14	12.00	7.44	7.10	5.40
MC-DM$^{\max}$	5.02	4.98	5.04	4.64	5.06	4.70	4.64	5.36	4.70	4.42	5.48	4.86
					$\theta = 0.9$ (two-steps-ahead forecasts)							
DM	21.92	12.32	7.94	6.74	23.10	12.54	7.60	6.22	22.14	11.62	8.40	6.28
HLN-DM	10.68	7.24	6.54	6.28	10.94	7.20	6.38	5.58	10.90	6.58	6.86	5.64
MC-DM$^{\max}$	5.02	4.84	4.42	4.76	4.84	5.34	4.78	5.10	4.44	4.90	4.88	4.64

Notes: Nominal level is 5 per cent. Results based on 5,000 replications.

Table 2. Empirical Test Power in Percentage

(a) $k = 0.75$												
	$\rho = 0$				$\rho = 0.5$				$\rho = 0.9$			
Test	T=8	T=16	T=32	T=64	T=8	T=16	T=32	T=64	T=8	T=16	T=32	T=64
					$\theta = 0$ (one-step-ahead forecasts)							
DM	5.60	7.42	11.6	18.28	6.70	8.32	15.04	25.04	9.68	18.70	39.94	70.24
MC-DM	5.82	7.64	11.36	19.00	6.06	8.38	14.18	24.56	10.50	20.02	39.28	69.00
					$\theta = 0.5$ (two-steps-ahead forecasts)							
DM	5.14	6.46	10.28	15.94	4.54	6.60	11.70	18.24	5.26	12.18	26.88	57.16
MC-DM$^{\max}$	4.88	6.36	7.88	13.26	5.82	6.30	9.14	14.60	6.70	11.32	22.68	45.90
					$\theta = 0.9$ (two-steps-ahead forecasts)							
DM	5.08	6.00	8.68	14.32	5.76	6.78	10.26	17.26	4.78	11.40	24.44	52.02
MC-DM$^{\max}$	5.16	5.62	7.82	11.96	5.26	6.46	9.16	14.98	5.64	11.68	20.98	42.86
(b) $k = 0.50$												
					$\theta = 0$ (one-step-ahead forecasts)							
DM	11.70	21.86	41.26	75.34	12.24	26.36	54.36	85.24	25.84	63.72	95.92	100
MC-DM	10.82	21.62	42.20	73.94	12.74	26.76	52.22	84.88	30.96	68.78	96.28	100
					$\theta = 0.5$ (two-steps-ahead forecasts)							
DM	4.92	12.94	30.86	62.06	5.04	15.50	36.94	73.12	5.00	35.80	85.74	99.80
MC-DM$^{\max}$	6.82	12.44	25.18	50.14	7.72	14.60	31.06	61.58	12.10	35.86	76.46	98.84
					$\theta = 0.9$ (two-steps-ahead forecasts)							
DM	5.82	11.52	29.00	56.60	4.92	15.50	35.64	69.48	6.26	36.50	80.32	99.24
MC-DM$^{\max}$	6.86	11.76	24.24	46.78	6.56	14.14	29.22	57.94	11.44	35.60	74.54	97.94

Notes: The DM test is compared with size-corrected critical values. Nominal level is 5 per cent. Results based on 5,000 replications.

Table 3. Canadian Term Structure of Interest Rates: Forecast Accuracy Results

Forecast period	Apr 94 - Mar 03		Apr 95 - Mar 03		Apr 96 - Mar 03		Apr 97 - Mar 03		Apr 98 - Mar 03	
Length of estimation window (months)	12		24		36		48		60	
Length of loss-differential series	108		96		84		72		60	
Loss function	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Full-sample tests										
DM	53.10	22.75	32.29	21.06	10.94	9.30	3.55*	1.91*	1.68*	1.06*
HLN-DM	54.18	24.11	33.81	22.57	12.42	10.69	4.61*	2.67*	2.56*	1.73*
MC-DM$^{\max}$	83.25	73.40	58.25	2.65*	11.60	4.60*	1.55*	0.15*	0.20*	0.30*
First subsample tests										
DM	82.27	41.96	70.52	92.54	68.96	76.82	33.66	41.95	3.40	7.16
HLN-DM	82.64	43.15	71.21	92.71	69.79	77.45	35.66	43.74	5.28	9.52
MC-DM	84.15	44.20	69.40	93.30	70.15	78.00	35.70	44.25	4.60	9.60
Second subsample tests										
DM	54.91	49.96	26.86	0.58*	5.08	1.52*	0.00*	0.00*	0.07*	0.02*
HLN-DM	55.85	51.00	28.46	1.08*	6.58	2.44	0.04*	0.02*	0.39*	0.17*
MC-DM	60.05	50.65	31.55	0.85*	5.85	2.40	0.05*	0.10*	0.15*	0.30*
Third subsample tests										
DM	40.81	29.09	75.98	53.60	34.37	16.26	64.78	35.01	76.45	15.26
HLN-DM	42.02	30.48	76.55	54.68	36.07	18.16	65.89	36.98	77.34	17.93
MC-DM	46.70	30.40	78.35	53.55	38.55	18.20	57.40	37.15	81.75	16.70

Note: Entries are two-sided p-values (in percentages) of tests of the null of equal forecast accuracy between the models in (20) and (21). An asterisk indicates a rejection of the null at the overall 5 per cent level for the corresponding column based on the decision rule: reject the null when the full-sample test p-value is less than 5 per cent; reject the null when a subsample test p-value is less than 5/3 per cent. Data are monthly 3-month and 6-month Canadian treasury bill rates from 1993 to 2003.

Table 4. U.S. Term Structure of Interest Rates: Forecast Accuracy Results

Forecast period										
	Apr 94 - Mar 03		Apr 95 - Mar 03		Apr 96 - Mar 03		Apr 97 - Mar 03		Apr 98 - Mar 03	
Length of estimation window (months)										
	12		24		36		48		60	
Length of loss-differential series										
	108		96		84		72		60	
Loss function										
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Full-sample tests										
DM	87.66	6.45	76.49	20.59	92.52	27.41	57.32	39.06	66.48	50.73
HLN-DM	87.97	7.38	77.14	22.10	92.76	29.17	58.83	41.00	67.95	52.77
MC-DM$^{\max}$	99.70	36.15	90.40	47.80	93.45	48.75	35.15	31.50	37.90	49.95
First subsample tests										
DM	54.31	29.13	86.07	21.20	55.59	95.06	45.34	74.05	43.54	90.96
HLN-DM	55.26	30.53	86.39	22.85	56.78	95.19	47.04	74.87	45.65	91.31
MC-DM	55.95	31.00	87.80	23.85	60.20	95.35	48.90	73.65	46.20	91.55
Second subsample tests										
DM	68.22	23.18	86.67	39.61	72.61	19.78	1.97	8.33	3.66	15.51
HLN-DM	68.88	24.64	86.99	40.99	73.35	21.68	3.20	10.35	5.58	18.19
MC-DM	99.65	26.50	98.60	41.45	79.85	20.30	3.00	11.40	5.30	18.50
Third subsample tests										
DM	4.59	5.65	21.44	16.94	61.30	49.98	83.29	80.88	99.64	94.93
HLN-DM	5.70	6.84	23.09	18.60	62.34	51.32	83.82	81.48	99.65	95.12
MC-DM	6.25	6.00	23.85	19.55	63.95	50.90	84.10	82.35	99.60	95.80

Notes: Entries are two-sided p-values (in percentages) of tests of the null of equal forecast accuracy between the models in (20) and (21). Data are monthly 3-month and 6-month U.S. treasury bill rates from 1993 to 2003.

Appendix

Table A1. Canadian 3-Month (TB3) and 6-Month (TB6) Treasury Bill Rates

Month	TB3	TB6	Month	TB3	TB6	Month	TB3	TB6
Mar 93	5.35	5.75	Aug 96	4.02	4.32	Jan 00	5.05	5.31
Apr 93	5.36	5.70	Sep 96	3.86	4.13	Feb 00	4.96	5.32
May 93	4.82	5.22	Oct 96	3.17	3.33	Mar 00	5.27	5.55
Jun 93	4.53	4.81	Nov 96	2.73	2.89	Apr 00	5.43	5.75
Jul 93	4.06	4.41	Dec 96	2.85	3.24	May 00	5.67	5.97
Aug 93	4.57	4.75	Jan 97	2.87	3.21	Jun 00	5.53	5.79
Sep 93	4.69	5.11	Feb 97	2.91	3.17	Jul 00	5.61	5.73
Oct 93	4.40	4.59	Mar 97	3.14	3.45	Aug 00	5.58	5.74
Nov 93	4.08	4.33	Apr 97	3.14	3.55	Sep 00	5.56	5.71
Dec 93	3.87	4.04	May 97	2.99	3.39	Oct 00	5.61	5.72
Jan 94	3.63	3.71	Jun 97	2.86	3.19	Nov 00	5.62	5.72
Feb 94	3.84	4.17	Jul 97	3.29	3.62	Dec 00	5.49	5.46
Mar 94	5.47	6.04	Aug 97	3.11	3.68	Jan 01	5.11	5.00
Apr 94	5.86	6.28	Sep 97	2.86	3.49	Feb 01	4.87	4.80
May 94	6.14	6.55	Oct 97	3.59	3.82	Mar 01	4.58	4.52
Jun 94	6.38	7.29	Nov 97	3.67	4.11	Apr 01	4.43	4.40
Jul 94	5.76	6.64	Dec 97	3.99	4.56	May 01	4.34	4.41
Aug 94	5.52	5.79	Jan 98	4.10	4.42	Jun 01	4.30	4.37
Sep 94	5.20	5.69	Feb 98	4.57	4.84	Jul 01	4.07	4.10
Oct 94	5.39	6.04	Mar 98	4.59	4.70	Aug 01	3.80	3.79
Nov 94	5.86	6.52	Apr 98	4.85	4.97	Sep 01	3.05	2.96
Dec 94	7.14	8.12	May 98	4.75	4.97	Oct 01	2.34	2.26
Jan 95	8.10	8.47	Jun 98	4.87	5.04	Nov 01	2.07	2.13
Feb 95	8.11	8.15	Jul 98	4.94	5.13	Dec 01	1.95	1.95
Mar 95	8.29	8.35	Aug 98	4.91	5.25	Jan 02	1.96	2.11
Apr 95	7.87	7.87	Sep 98	4.91	5.03	Feb 02	2.05	2.19
May 95	7.40	7.36	Oct 98	4.71	4.73	Mar 02	2.30	2.68
Jun 95	6.73	6.65	Nov 98	4.78	4.88	Apr 02	2.37	2.68
Jul 95	6.65	6.87	Dec 98	4.66	4.76	May 02	2.60	2.87
Aug 95	6.34	6.62	Jan 99	4.68	4.76	Jun 02	2.70	2.87
Sep 95	6.58	6.80	Feb 99	4.87	4.97	Jul 02	2.81	2.90
Oct 95	7.16	7.21	Mar 99	4.63	4.73	Aug 02	2.96	3.08
Nov 95	5.83	5.87	Apr 99	4.60	4.66	Sep 02	2.83	2.93
Dec 95	5.54	5.64	May 99	4.48	4.71	Oct 02	2.73	2.81
Jan 96	5.12	5.20	Jun 99	4.56	4.77	Nov 02	2.71	2.81
Feb 96	5.21	5.38	Jul 99	4.71	4.82	Dec 02	2.63	2.75
Mar 96	5.02	5.25	Aug 99	4.68	4.87	Jan 03	2.83	2.99
Apr 96	4.78	4.97	Sep 99	4.66	4.87	Feb 03	2.88	3.06
May 96	4.68	4.88	Oct 99	4.87	5.19	Mar 03	3.14	3.34
Jun 96	4.70	4.94	Nov 99	4.73	4.96			
Jul 96	4.39	4.75	Dec 99	4.85	5.16			

Table A2. U.S. 3-Month (TB3) and 6-Month (TB6) Treasury Bill Rates

Month	TB3	TB6	Month	TB3	TB6	Month	TB3	TB6
Mar 93	2.95	3.05	Aug 96	5.05	5.13	Jan 00	5.32	5.50
Apr 93	2.87	2.97	Sep 96	5.09	5.24	Feb 00	5.55	5.72
May 93	2.96	3.07	Oct 96	4.99	5.11	Mar 00	5.69	5.85
Jun 93	3.07	3.20	Nov 96	5.03	5.07	Apr 00	5.66	5.81
Jul 93	3.04	3.16	Dec 96	4.91	5.04	May 00	5.79	6.10
Aug 93	3.02	3.14	Jan 97	5.03	5.10	Jun 00	5.69	5.97
Sep 93	2.95	3.06	Feb 97	5.01	5.06	Jul 00	5.96	6.00
Oct 93	3.02	3.12	Mar 97	5.14	5.26	Aug 00	6.09	6.07
Nov 93	3.10	3.26	Apr 97	5.16	5.37	Sep 00	6.00	5.98
Dec 93	3.06	3.23	May 97	5.05	5.30	Oct 00	6.11	6.04
Jan 94	2.98	3.15	Jun 97	4.93	5.13	Nov 00	6.17	6.06
Feb 94	3.25	3.43	Jul 97	5.05	5.12	Dec 00	5.77	5.68
Mar 94	3.50	3.78	Aug 97	5.14	5.19	Jan 01	5.15	4.95
Apr 94	3.68	4.09	Sep 97	4.95	5.09	Feb 01	4.88	4.71
May 94	4.14	4.60	Oct 97	4.97	5.09	Mar 01	4.42	4.28
Jun 94	4.14	4.55	Nov 97	5.14	5.17	Apr 01	3.87	3.85
Jul 94	4.33	4.75	Dec 97	5.16	5.24	May 01	3.62	3.62
Aug 94	4.48	4.88	Jan 98	5.04	5.03	Jun 01	3.49	3.45
Sep 94	4.62	5.04	Feb 98	5.09	5.07	Jul 01	3.51	3.45
Oct 94	4.95	5.39	Mar 98	5.03	5.04	Aug 01	3.36	3.29
Nov 94	5.29	5.72	Apr 98	4.95	5.06	Sep 01	2.64	2.63
Dec 94	5.60	6.21	May 98	5.00	5.14	Oct 01	2.16	2.12
Jan 95	5.71	6.21	Jun 98	4.98	5.12	Nov 01	1.87	1.88
Feb 95	5.77	6.03	Jul 98	4.96	5.03	Dec 01	1.69	1.78
Mar 95	5.73	5.89	Aug 98	4.90	4.95	Jan 02	1.65	1.73
Apr 95	5.65	5.77	Sep 98	4.61	4.63	Feb 02	1.73	1.82
May 95	5.67	5.67	Oct 98	3.96	4.05	Mar 02	1.79	2.01
Jun 95	5.47	5.42	Nov 98	4.41	4.42	Apr 02	1.72	1.93
Jul 95	5.42	5.37	Dec 98	4.39	4.40	May 02	1.73	1.86
Aug 95	5.40	5.41	Jan 99	4.34	4.33	Jun 02	1.70	1.79
Sep 95	5.28	5.30	Feb 99	4.44	4.44	Jul 02	1.68	1.70
Oct 95	5.28	5.32	Mar 99	4.44	4.47	Aug 02	1.62	1.60
Nov 95	5.36	5.27	Apr 99	4.29	4.37	Sep 02	1.63	1.60
Dec 95	5.14	5.13	May 99	4.50	4.56	Oct 02	1.58	1.56
Jan 96	5.00	4.92	Jun 99	4.57	4.82	Nov 02	1.23	1.27
Feb 96	4.83	4.77	Jul 99	4.55	4.58	Dec 02	1.19	1.24
Mar 96	4.96	4.96	Aug 99	4.72	4.87	Jan 03	1.17	1.20
Apr 96	4.95	5.06	Sep 99	4.68	4.88	Feb 03	1.17	1.18
May 96	5.02	5.12	Oct 99	4.86	4.98	Mar 03	1.13	1.13
Jun 96	5.09	5.25	Nov 99	5.07	5.20			
Jul 96	5.15	5.30	Dec 99	5.20	5.44			

Bank of Canada Working Papers
Documents de travail de la Banque du Canada

Working papers are generally published in the language of the author, with an abstract in both official languages. *Les documents de travail sont publiés généralement dans la langue utilisée par les auteurs; ils sont cependant précédés d'un résumé bilingue*.

2004

2004-1	The Effect of Adjustment Costs and Organizational Change on Productivity in Canada: Evidence from Aggregate Data	D. Leung

2003

2003-44	Common Trends and Common Cycles in Canadian Sectoral Output	F. Barillas and C. Schleicher
2003-43	Why Does Private Consumption Rise After a Government Spending Shock?	H. Bouakez and N. Rebei
2003-42	A Structural VAR Approach to the Intertemporal Model of the Current Account	T. Kano
2003-41	Anatomy of a Twin Crisis	R.H. Solomon
2003-40	Poignée de main invisible et persistance des cycles économiques : une revue de la littérature	C. Calmès
2003-39	Alternative Targeting Regimes, Transmission Lags, and the Exchange Rate Channel	J.-P. Lam
2003-38	Simple Monetary Policy Rules in an Open-Economy, Limited-Participation Model	S. Hendry, W-M. Ho, and K. Moran
2003-37	Financial Constraints and Investment: Assessing the Impact of a World Bank Loan Program on Small and Medium-Sized Enterprises in Sri Lanka	V. Aivazian, D. Mazumdar, and E. Santor
2003-36	Excess Collateral in the LVTS: How Much is Too Much?	K. McPhail and A. Vakos
2003-35	Real Exchange Rate Persistence in Dynamic General-Equilibrium Sticky-Price Models: An Analytical Characterization	H. Bouakez
2003-34	Governance and Financial Fragility: Evidence from a Cross-Section of Countries	M. Francis
2003-33	Do Peer Group Members Outperform Individual Borrowers? A Test of Peer Group Lending Using Canadian Micro-Credit Data	R. Gomez and E. Santor

Copies and a complete list of working papers are available from:
Pour obtenir des exemplaires et une liste complète des documents de travail, prière de s'adresser à :

Publications Distribution, Bank of Canada
234 Wellington Street, Ottawa, Ontario K1A 0G9
E-mail: publications@bankofcanada.ca
Web site: http://www.bankofcanada.ca

Diffusion des publications, Banque du Canada
234, rue Wellington, Ottawa (Ontario) K1A 0G9
Adresse électronique : publications@banqueducanada.ca
Site Web : http://www.banqueducanada.ca