On Testing Equal Conditional Predictive Ability Under Measurement Error

Yannick Hoga and Timo Dimitriadis

Faculty of Economics and Business Administration, University of Duisburg-Essen, Essen, Germany; Alfred Weber Institute of Economics, Heidelberg University, Heidelberg, Germany; Heidelberg Institute for Theoretical Studies, Heidelberg, Germany

ABSTRACT

Loss functions are widely used to compare several competing forecasts. However, forecast comparisons are often based on mismeasured proxy variables for the true target. We introduce the concept of exact robustness to measurement error for loss functions and fully characterize this class of loss functions as the Bregman class. Hence, only conditional mean forecasts can be evaluated exactly robustly. For such exactly robust loss functions, forecast loss differences are on average unaffected by the use of proxy variables and, thus, inference on conditional predictive ability can be carried out as usual. Moreover, we show that more precise proxies give predictive ability tests higher power in discriminating between competing forecasts. Simulations illustrate the different behavior of exactly robust and nonrobust loss functions. An empirical application to U.S. GDP growth rates demonstrates the nonrobustness of quantile forecasts. It also shows that it is easier to discriminate between mean forecasts issued at different horizons if a better proxy for GDP growth is used.

1. Motivation

Due to the central role of forecasts in economic policy, business, climate research and beyond, forecast comparisons have a long tradition. Such comparisons rely on a (statistically or economically motivated) loss function that measures the loss as a function of the issued forecast and the realization of the target variable. Since the seminal contribution of Diebold and Mariano (1995), tests of equal predictive ability (EPA) have played a central role in comparing competing forecasts. Giacomini and White (2006) extended EPA tests by introducing tests of equal conditional predictive ability (ECPA), where the conditioning is, for example, on current economic conditions. The null hypothesis of ECPA tests is that the conditional means of the forecast losses are identical.

E(C)PA tests are studied extensively under estimation error in the forecasts (see, e.g., West 1996; Clark and McCracken 2001; Patton 2020). However, the effect of measurement error in the observed target variable has not received as much attention. Some exceptions—to be discussed below—are the works of Hansen and Lunde (2006), Patton (2011), Laurent, Rombouts, and Violante (2013) and Li and Patton (2018). Nonetheless, measurement error is present in many economic and financial time series. Examples in economics include the gross domestic product (GDP) (Aruoba et al. 2016), inflation rates (Clark and McCracken 2009; Fox and Syed 2016), and job earnings (Abowd and Stinson 2013). In finance, the conditional variance of asset returns can only be approximated by the squared return or by high-frequency measures such as realized volatility (Andersen et al. 2013). Examples beyond economics and finance include, among others, meteorological applications such as the measurement of precipitation or wind speeds (Ferro 2017).

As a consequence, many forecast comparisons are carried out with approximated, mismeasured target variables, also called proxies. In such a case, the forecast losses—as measured by some loss function—may be systematically different from those obtained using the actual, but latent, target variable. Hence, differences in predictive ability may be clouded by the use of such proxies. In this article, we derive conditions under which ECPA tests can be validly carried out if only some proxy for the target variable is available. Of course, the proxy needs to resemble the target variable in some way. As a minimal “resemblance condition”, we require it to be conditionally unbiased for the true target. If several alternative (conditionally unbiased) proxies are available, then we further derive conditions which proxy entails the most powerful tests.

To do so, we define a loss function to be exactly robust to measurement error if the (conditional) expectation of the forecast loss differences is unchanged when using the proxy instead of the true target variable. Since most of the literature on forecast evaluation is concerned with univariate quantities (Gneiting 2011; Patton 2011), it is worth stressing that the target variable and the forecasts may be multivariate here. Some work on characterizing strictly consistent loss functions for the specific multivariate mean functional can be found in Banerjee, Guo, and Wang (2005), Laurent, Rombouts, and Violante (2013), and Frongillo and Kash (2015).

CONTACT Yannick Hoga yannick.hoga@vwl.uni-due.de Universitat Duisburg-Essen - Campus Essen, Faculty of Economics, Universitätsstraße 12, Essen, 45141 Germany.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/UBES.
Our first main contribution is to characterize the loss functions that are exactly robust to measurement error. We show that the class of exactly robust loss functions coincides with the Bregman loss functions (Banerjee, Guo, and Wang 2005, def. 1), and the class to which Patton (2011) referred as “robust” loss functions in the univariate case. This implies that only conditional mean forecasts can be compared robustly in ECPA tests. While this is of course rather restrictive, for many economic variables that are measured with error, the conditional mean is precisely the object of interest; for example, the conditional mean of GDP growth or inflation, or the conditional variance of asset returns (which commonly coincides with the conditional mean of squared returns). The importance of the conditional mean is further underscored by the prevalence in economics of (V)ARMA-type models, which are designed to dynamically model the conditional mean.

The studies most closely related to our first contribution are those of Patton (2011) (in the univariate case) and Laurent, Rombouts, and Violante (2013) (in the multivariate case), both of which are devoted to the specific task of comparing conditional variance forecasts for financial returns using high-frequency (HF) proxies. Building on Hansen and Lunde (2006), Patton (2011) characterized loss functions that give consistent relative rankings when only a conditionally unbiased proxy for the conditional variance is available. (His work was later on extended by Laurent, Rombouts, and Violante (2013) to conditional variance-covariance matrix forecasts.) Our exact robustness is a stronger requirement than Patton’s (2011) ordering robustness. While under exact robustness two forecasts have the same conditional predictive ability for the true target and the proxy, ordering robustness merely implies that the ordering of the two forecasts is preserved when a proxy is used instead of the true target; however, the magnitude in predictive ability may be changed. For instance, the average forecast loss differences may be large for the true target, yet very small when the proxy is used. Thus, while the ranking (in population) is preserved, it may be much harder to discriminate between the two forecasts in finite samples when only the proxy is available. In contrast, under exact robustness, the magnitude of the expected differences is identical for the true target and the proxy. Thus, our concept of exact robustness almost immediately implies that ECPA tests can be carried out as usual under measurement error, in particular allowing us to do a local power analysis. While our exact robustness is a stronger requirement than Patton’s (2011) ordering robustness, we show in Theorem 1 that—surprisingly—the respective classes of loss functions coincide. By doing so, we also refine the results of Patton (2011) and Laurent, Rombouts, and Violante (2013) along several dimensions; see Remarks 3 and 4 for details.

An important aspect of our first main contribution is that—unlike Patton (2011) and Laurent, Rombouts, and Violante (2013)—we do not restrict attention to the mean functional from the outset. Thus, by narrowing down the class of functionals that can be evaluated robustly to the mean functional, we are able to show that, for example, the ranking of median forecasts may be affected by noisy proxies. However, the absolute error loss—pertaining to median forecasts—has regularly been used in comparing forecasts for mismeasured variables, such as inflation (Hansen 2005; Medeiros et al. 2021), GDP growth (Rudebusch and Williams 2009; Bräuning and Koopman 2014), and integrated variances (Hansen and Lunde 2005). Likewise, quantile forecasts have been compared for GDP growth (Brownlees and Souza 2021). In each case, our results suggest that these comparisons should be interpreted with caution, due to the nonrobustness of quantiles. We view this implication of the present work to have the strongest impact on empirical practice, because it narrows down the class of functionals which can be evaluated validly under measurement error to the mean functional.

In Appendix A, we strengthen our nonrobustness results by showing that median (and more generally, quantile) forecasts cannot be evaluated exactly robustly, even when the conditional unbiasedness assumption is replaced by any other “resemblance condition” on the proxy. In other words, an exactly robust evaluation of quantile forecasts requires the proxy and the true target to coincide, that is, it requires the absence of measurement error.

Our second main contribution is to study the local power of ECPA tests using proxy variables and exactly robust loss functions. We demonstrate that power increases for more accurate proxies. The (infeasible) upper bound for the test power is obtained when evaluating forecasts with the most accurate proxy. In our case, this “proxy” is the—generally even ex post—latent forecast target, that is, the conditional mean of the target variable. Our results lend theoretical support to the intuition that it is easier to discriminate between competing forecasts if the target is approximated more precisely.

Our simulations show that for mean forecasts, the asymptotic local power of the proxy-based ECPA test provides a good approximation in finite samples. We further demonstrate the dangers of evaluating forecasts for other functionals than the mean (and hence, of using nonrobust loss functions) for comparing predictive accuracy with proxy variables: We demonstrate that the expected loss difference of quantile forecasts can change its sign if a conditionally unbiased proxy is used instead of the true target variable. Appendix D further shows that size distortions may arise and, for certain alternatives, a loss of power occurs. These drawbacks, instead of getting less serious in larger samples, get more pronounced as the sample size increases.

We apply our proxy-based ECPA test to GDP growth rates in the United States. GDP is a measure of the aggregate real output of the economy and, as such, is perhaps the most important macroeconomic indicator. However, GDP (and, hence, also GDP growth) cannot be measured exactly for various reasons. For instance, relevant economic census data are collected only once every five years and tax returns are incorporated into the national accounts only over time, leading to frequent revisions of GDP estimates.

Several proxies for true GDP growth (denoted ΔGDP) are available (Landefeld, Seskin, and Fraumeni 2008). The arguably most popular proxy is the expenditure-side approximation ΔGDP_e, followed by the income-side proxy ΔGDP_i. Our third proxy, ΔGDP_p, from Aruoba et al. (2016), combines both of these information sources as a Kalman-smoothed extraction from a dynamic factor model for ΔGDP_e and ΔGDP_i. Since Aruoba et al. (2016, p. 393) “advocate the additional calculation of ΔGDP_p and using it as the benchmark ΔGDP estimate,” we expect ΔGDP_p to be the most precise proxy of latent growth.
We first show that the ranking of mean forecasts obtained from autoregressive models and the Survey of Professional Forecasters (SPF) is consistent among these proxies. In contrast, the loss differences can vary substantially for GDP growth quantile forecasts, which recently gained attention in the Growth-at-Risk (GaR) literature (Adrian, Boyarchenko, and Giannone 2019; Adrian et al. 2021). Second, our theory suggests that using \( \Delta \text{GDP}_+ \) in ECPA tests leads to better discrimination between different forecasts. This is indeed what we find when comparing mean predictions issued at different horizons. Naturally, we expect \( \Delta \text{GDP}_+ \) forecasts, which recently gained attention in the Growth-at-Risk (GaR) literature (Adrian, Boyarchenko, and Giannone 2019; SPF) is consistent among these proxies. In contrast, the leading case \( \tau = 1 \) give futher details for the empirical application.

The remainder of the article proceeds as follows. In Section 2, we define exact/ordering robustness to measurement error for loss functions, characterize these loss functions and show that exact and ordering robustness are equivalent. Then, we derive the local power of ECPA tests based on proxy variables and robust loss functions. Section 3 numerically illustrates the robustness and local power results in Monte Carlo experiments. Section 4 applies our test to U.S. GDP growth forecasts. Finally, Section 5 concludes. The supplementary material establishes the nonrobustness of quantile forecasts under any “resemblance condition” on the proxy, and it contain all proofs and technical derivations. It also provides additional simulation results and gives further details for the empirical application.

2. Main Results

2.1. Characterizing Robust Loss Functions

Let \( (\Omega, \mathcal{A}, P) \) be a probability space. If not stated otherwise, then all (in-)equalities involving conditional expectations are tacitly assumed to hold \( P \)-almost surely (a.s.) in the following. Let \( \mathcal{P} \) denote some class of distribution functions on \( \mathbb{R}^k \) (\( k \in \mathbb{N} \)), which is specified later on. For some time point \( t \in \mathbb{N} \), we denote the target variable (e.g., GDP growth) by \( Y_t: \Omega \rightarrow \mathbb{R} \subset \mathbb{R}^k \), the time-(\( t-1 \)) information set by \( \mathcal{F}_{t-1} \), and the conditional distribution of \( Y_t \) given \( \mathcal{F}_{t-1} \) by \( F_t(\omega; \cdot) = P \{ Y_t \leq \cdot \mid \mathcal{F}_{t-1}(\omega) \} \), where we assume that \( F_t(\omega; \cdot) \in \mathcal{P} \) for all \( \omega \in \Omega \). We denote by \( F_t(\cdot) \) the random variable defined by the mapping \( \omega \mapsto F_t(\omega; \cdot) \).

Remark 1. The theory of this article can be extended to \( \tau \)-step ahead forecasts for \( \tau \geq 2 \) in a straightforward fashion by considering the information set \( \mathcal{F}_{t-\tau} \) instead of \( \mathcal{F}_{t-1} \). Since we leave \( \mathcal{F}_{t-1} \) unspecified in the following, this can be seen by letting \( \mathcal{F}_{t-1} \) contain only information available at time \( (t-\tau) \). We deliberately do not make this explicit in the notation in order to keep the exposition as simple as possible.

The aim in a forecasting situation is to predict some target functional \( T: \mathcal{P} \rightarrow \mathbb{R}^k \) of \( F_t \). We denote the resulting forecast target by \( x_t^\ast: \Omega \rightarrow A, \omega \mapsto T(F_t(\omega; \cdot)) \). For the leading case \( k = 1 \), this may be the conditional mean of GDP growth, \( x_t^\ast = E_{t-1}[Y_t] \), where we write \( E_{t-1}[\cdot] = E[\cdot \mid \mathcal{F}_{t-1}] \) for short. We assume that there exists a strictly \( \mathcal{P} \)-consistent (or simply strictly consistent) loss function \( L: \Omega \times A \rightarrow \mathbb{R} \) for the functional \( T \), that is,

\[
E_{Y \sim F}[L(Y, T(F))] \leq E_{Y \sim F}[L(Y, x)],
\]

for all \( F \in \mathcal{P} \) and all \( x \in A \), and equality in (1) implies \( x = T(F) \); see Gneiting (2011). Here, the notation \( E_{Y \sim F}[\cdot] \) denotes the expectation with respect to \( Y \) with distribution \( F \). A functional for which a strictly consistent loss function exists is termed elicitable. Assuming \( T \) to be elicitable is not restrictive in our context, because when no strictly consistent scoring function exists, forecasts cannot be compared validly (Gneiting 2011). Denote by \( x_{it} \mid \Omega \rightarrow A (i = 1, 2) \) the two competing, \( \mathcal{F}_{t-1} \)-measurable forecasts of \( x_t^\ast \). The forecast loss difference

\[
d(Y_{it}, x_{1it}, x_{2it}) = L(Y_{it}, x_{1it}) - L(Y_{it}, x_{2it})
\]

measures the relative performance of \( x_{1it} \) and \( x_{2it} \). Since the loss function is negatively oriented, a negative (positive) loss difference favors \( x_{1it} \) (\( x_{2it} \)).

As pointed out in the motivation, the true \( Y_t \) may often not be available for comparing forecasts due to measurement error. Instead, one has to rely on a proxy \( \tilde{Y}_t: \Omega \rightarrow \mathbb{R} \) for \( Y_t \) and use \( d(\tilde{Y}_t, x_{1t}, x_{2t}) \). Of course, the proxy has to bear some resemblance to the target. To ensure this, we assume that \( \tilde{Y}_t \) is conditionally unbiased for \( Y_t \), that is, that \( E_{t-1}[\tilde{Y}_t] = E_{t-1}[Y_t] \).

This also implies that (mean) differences between \( Y_t \) and \( \tilde{Y}_t \) cannot be predicted from information available at time \( t-1 \). We consider this to be a natural assumption for a proxy variable, since otherwise one could predict the average measurement error \( \tilde{Y}_t - Y_t \) based on \( \mathcal{F}_{t-1} \). We refer to the empirical application for a modeling framework of GDP growth rates, where the conditional unbiasedness assumption is satisfied for \( Y_t = \Delta \text{GDP}_t \) and \( \tilde{Y}_t \in \{ \Delta \text{GDP}_t, \Delta \text{GDP}_{t-1}, \Delta \text{GDP}_{t+1} \} \); see in particular (16).

We denote the conditional distribution function of \( \tilde{Y}_t \mid \mathcal{F}_{t-1} \) by \( F_t(\omega; \cdot) \) for all \( \omega \in \Omega \) and the corresponding random variable by \( \tilde{F}_t(\cdot) \). The following definition ensures that the expected forecast loss differences are the same for \( Y_t \) and a conditionally unbiased proxy \( \tilde{Y}_t \).

Definition 1. \( L(\cdot, \cdot) \) is exactly robust to measurement error (or simply: exactly robust) with respect to \( \mathcal{P} \), if

\[
E_{t-1}[d(Y_{it}, x_{1it}, x_{2it})] = E_{t-1}[d(\tilde{Y}_{it}, x_{1it}, x_{2it})] \quad \text{a.s.}
\]

for all \( \mathcal{F}_{t-1} \)-measurable forecasts \( x_{1t} \) and \( x_{2t} \), and all \( Y_t \) and all conditionally unbiased proxies \( \tilde{Y}_t \) with \( F_t(\omega; \cdot) \in \mathcal{P} \) and \( \tilde{F}_t(\omega; \cdot) \in \mathcal{P} \) for all \( \omega \in \Omega \).

Exact robustness to measurement error suggests that we can learn as much about the relative merits of \( x_{1t} \) and \( x_{2t} \) by observing \( d(\tilde{Y}_{it}, x_{1it}, x_{2it}) \) instead of \( d(Y_{it}, x_{1it}, x_{2it}) \). However, this is only correct in expectation, or equivalently, asymptotically under a suitable law of large numbers. In finite samples, this is unfortunately not true, since \( d(\tilde{Y}_{it}, x_{1it}, x_{2it}) \) may have a larger (or smaller) variance than \( d(Y_{it}, x_{1it}, x_{2it}) \).

The more restricted the class \( \mathcal{P} \) in Definition 1, the richer the class of exactly robust loss functions. An extreme case arises if \( \mathcal{P} = \{ F : \mathbb{R}^k \rightarrow [0, 1] \mid F(x) = \mathbf{1}_{\{x \geq o\}} \text{ for some } o \in \mathbb{O} \} \) is the class of degenerate distributions. (The inequality \( x \geq o \)) is to be understood componentwise if the quantities involved...
are vector-valued.) Then, by (a.s.) constancy of $Y_t$ and $\hat{Y}_t$, we necessarily have that $Y_t \approx_{\text{a.s.}} \hat{Y}_t$ by conditional unbiasedness. The latter implies that any loss function is exactly robust with respect to this rather restricted class $\mathcal{P}$.

**Remark 2.** Consider univariate log-returns $r_t$ on some speculative asset and assume that $E[r_t | \mathcal{F}_{t-1}] = 0$, as is common for financial data. In this setting, forecasts of the (latent) conditional variance $x_t^* = E[r_t^2 | \mathcal{F}_{t-1}] = \text{var}(r_t | \mathcal{F}_{t-1})$ are essential for risk management purposes (Andersen et al. 2013). The related literature often does not make the distinction between the target variable and the forecast target explicit (Hansen and Lunde 2006; Patton 2011; Laurent, Rombouts, and Violante 2013). This is however crucial for us as we operate under the decision-theoretic setting of Gneiting (2011). Doing so allows us to narrow down the target functionals which can be evaluated exactly robustly to the mean functional in **Theorem 1**.

Thus, in our setting, the target variable is $Y_t = r_t^2$ and the target functional $T$ is the mean. To compare two volatility forecasts $x_{1t}$ and $x_{2t}$, the natural choice is then $Y_t = r_t^2$. However, in the—different, but related—context of evaluating volatility forecasts of GARCH models, Andersen and Bollerslev (1998) advocated the use of less noisy high-frequency proxies $\tilde{Y}_t$ of the conditional variance (such as realized volatility or range-based volatility) satisfying $E_{t-1}[\tilde{Y}_t] = E_{t-1}[Y_t] = x_t^*$. This example shows that (other than our notation for $Y_t$ and $\hat{Y}_t$ suggests) $\hat{Y}_t$ does not necessarily have to be regarded as coming "close" to $Y_t$. Instead, we may sometimes interpret $Y_t$ and $\hat{Y}_t$ as providing two differently conditionally unbiased estimates of the forecast target $x_t^*$.

**Remark 3.** In the framework of **Remark 2**, Patton (2011) investigated conditions under which loss functions produce consistent relative rankings in the sense that

$$E[L(Y_t, x_{1t})] \leq E[L(Y_t, x_{2t})]$$

$$\implies E[L(\hat{Y}_t, x_{1t})] \leq E[L(\hat{Y}_t, x_{2t})]$$

for any (volatility) proxy satisfying $E_{t-1}[\hat{Y}_t] = E_{t-1}[Y_t] (= x_t^*)$. This property is conceptually different from exact robustness to measurement error in that, first, only the ranking of forecasts is concerned and, second, unconditional expectations are considered. Laurent, Rombouts, and Violante (2013) generalized Patton’s (2011) results by considering multivariate $r_t$, where $Y_t = \text{vech}(r_t r_t')$ is the target variable and the forecast target $x_t^* = E[Y_t | \mathcal{F}_{t-1}]$ is the (vech-transformed) conditional variance-covariance matrix of the returns. Here, vech$(\cdot)$ stacks the lower triangular part of a matrix into a vector.

Despite the conceptual differences, it will be insightful to transfer the idea of loss functions that produce consistent relative rankings to our framework. To do so, we generalize the definition of Patton (2011) as follows:

**Definition 2.** $L(\cdot, \cdot)$ is ordering robust to measurement error (or simply: ordering robust) with respect to $\mathcal{P}$, if

$$E_{t-1}[d(Y_t, x_{1t}, x_{2t})] \leq 0 \quad \text{a.s.}$$

$$\implies E_{t-1}[d(\hat{Y}_t, x_{1t}, x_{2t})] \leq 0 \quad \text{a.s.}$$

for all $\mathcal{F}_{t-1}$-measurable forecasts $x_{1t}$ and $x_{2t}$, and all $Y_t$ and all conditionally unbiased proxies $\hat{Y}_t$ with $F_t(\omega, \cdot) \in \mathcal{P}$ and $\hat{F}_t(\omega, \cdot) \in \mathcal{P}$ for all $\omega \in \Omega$.

At first sight, exact robustness seems much stronger than ordering robustness: For the expected loss differences, the former requires equality while the latter is already satisfied by matching signs. Nonetheless, foreshadowing one of the main results of **Theorem 1**, we show that the classes of exactly and ordering robust loss functions coincide under mild conditions. However, the concept of exact robustness opens up analyses, which are not possible based on ordering robustness only, such as our local power results for ECPA tests in Section 2.2. Thus, we argue that exact robustness is the theoretically more appealing notion.

Before we can state **Theorem 1**, recall the concept of a subgradient from convex analysis. To do so, we let $(\cdot, \cdot)$ denote the standard scalar product in $\mathbb{R}^k$. Then, the subgradient at value $x \in \mathcal{A}$ of some convex function $\phi : \mathcal{A} \to \mathbb{R}$, denoted $\partial \phi(x)$, is any vector $s \in \mathcal{A}$ satisfying $\phi(y) \geq \phi(x) + \langle s, y - x \rangle$ for all $y \in \mathcal{A}$ (Hiriart-Urrut and Lemaréchal 2001, def. 1.2.1). Recall from Hiriart-Urrut and Lemaréchal 2001 (sec. D) that such a vector always exists.

**Theorem 1.** Let $\mathcal{P}$ be a convex set of distribution functions. Assume that $T : \mathcal{P} \to \mathcal{A}$ is surjective, $\mathcal{A}$ is convex, and $L(\cdot, \cdot)$ is strictly $\mathcal{P}$-consistent for $T$. Then, the following are equivalent:

(a) $L(\cdot, \cdot)$ is of the form

$$L(Y, x) = \phi(x) + \langle d\phi(x), Y - x \rangle + a(Y),$$

where $\phi : \mathcal{A} \to \mathbb{R}$ is strictly convex with subgradient $d\phi(\cdot)$, and $a : \mathcal{O} \to \mathbb{R}$ is integrable with respect to all $F \in \mathcal{P}$;

(b) $L(\cdot, \cdot)$ is exactly robust with respect to $\mathcal{P}$;

(c) $L(\cdot, \cdot)$ is ordering robust with respect to $\mathcal{P}$;

(d) For all $Y_t$ with $F_t(\omega, \cdot) \in \mathcal{P}$ for all $\omega \in \Omega$, it holds that $x_t^* = T(F_t(\cdot)) \approx_{\text{a.s.}} E_{t-1}[Y_t]$.

**Theorem 1** offers three main insights. First, it characterizes the exactly robust loss functions, while making no assumption on the functional to be evaluated in advance. Exact robustness is essential to theoretically study ECPA tests under the alternative, where there is a difference in predictive ability, that is, $E_{t-1}[d(Y_t, x_{1t}, x_{2t})] \neq 0$. In the absence of the equivalence of (b) and (c) established in **Theorem 1**, mere ordering robustness would not be sufficient to do so, because the magnitude of the deviation from zero in $E_{t-1}[d(\tilde{Y}_t, x_{1t}, x_{2t})] \neq 0$ may be changed. Thus, a second contribution of **Theorem 1** is the equivalence of exact and ordering robustness. The third insight is that if, in practice, there is measurement error in the target variable, then only conditional mean forecasts can be compared (exact and/or ordering) robustly. For other functionals, such as the median, the forecast ranking is affected by the use of proxies. In particular, many commonly used loss functions, such as absolute error (AE) loss $L(Y, x) = |Y - x|$, are neither exactly robust nor ordering robust. Nonetheless, the AE loss has been used in comparing forecasts for misspecified variables, such as inflation (Hansen 2005; Medeiros et al. 2021), GDP growth (Rudebusch ...
and Williams 2009; Bräuning and Koopman 2014), and integrated variances (Hansen and Lunde 2005). Thus, Theorem 1
casts doubt on the rankings obtained by these comparisons.

Theorem 1 crucially depends on the conditional unbiasedness condition, \( E_{t-1}[\hat{Y}_t] = E_{t-1}[Y_t] \). This raises the question if forecasts for other target functionals than the mean can be compared robustly under alternative “resemblance conditions” on \( Y_t \) and \( \hat{Y}_t \). In Appendix A, we show that conditional quantile forecasts cannot be evaluated exactly robustly, unless one imposes the degenerate resemblance condition that the distributions of \( Y_t \) and \( \hat{Y}_t \) coincide. This reinforces our interpretation of the mean being the only target functional that allows for exactly robust evaluation. Thus, we have the surprising result that while the mean is more robust than the median in forecast comparisons, the opposite is well known to hold in classical estimation theory. We stress though that in estimation theory the robustness objective is different from that in the present article. While for the former the goal is to reduce the influence of outliers on estimation, our emphasis is on robustness of forecast comparisons to estimation error in the target variable.

**Remark 4.** The equivalence of (a) and (c) in Theorem 1 may be viewed as a generalization of Patton (2011, prop. 1) for \( k = 1 \) and of Laurent, Rombouts, and Violante (2013, prop. 2) for \( k \in \mathbb{N} \). Since Laurent, Rombouts, and Violante (2013) used very similar regularity conditions as Patton (2011), we only highlight the main improvements on the latter. First, while Patton’s characterization builds on the unconditional expectation, we consider conditional expectations, which is crucial for tests of equal conditional predictive ability (Giacomini and White 2006) and tests of superior conditional predictive ability (Li, Liao, and Quaedvlieg 2021). Second, similar to the property of strict consistency for loss functions (Gneiting 2011), robustness should be considered with respect to a specified class \( \mathcal{P} \) of distributions. While we merely require \( \mathcal{P} \) to be convex, Patton (2011, A2) restricted attention to absolutely continuous distributions. Third, Patton (2011) only considered continuously differentiable losses, which ignores important classes of loss functions, such as the generalized piecewise linear (GPL) losses. In contrast, our Theorem 1 dispenses with any regularity conditions on the class of possible loss functions. Thus, our class of loss functions in (2) is broader than his and we do not rule out nondifferentiable losses from the outset. Fourth, by considering \( x^2_t = E[r^2_t | \mathcal{F}_{t-1}] \) (cf. Remark 2), (Patton 2011, A1 and A4) specifies \( T \) to be the mean functional as an assumption, such that implicitly only Bregman loss functions are considered at the outset. In contrast, we do not specify \( T \) in advance, allowing us to show the nonrobustness of any functional apart from the mean. We view this final improvement as the most important one for empirical practice: when there is measurement error (such that the proxy is conditionally unbiased), robust conclusions on predictive ability are only possible for mean forecasts. In particular, the forecast ranking of any other functional (e.g., the median or quantiles) may be affected; see Sections 3.2 and 4.2 for simulations and applications for quantile forecasts.

**Remark 5.** The classification of loss functions in Theorem 1 can be employed (by invoking Theorem 2.5 of Dimitriadis, Fissler, and Ziegel 2021a) for the objective functions in M-estimation of semiparametric models, when only a proxy \( \hat{Y}_t \) of the response variable, \( Y_t \), is observable. Theorem 1 then implies that for conditionally unbiased proxies, the M-estimator is consistent for, and only for, conditional mean models; in contrast to, for example, conditional quantile models.

### 2.2. Testing Equal Predictive Accuracy

Section 2.1 shows that for loss functions of the form (2), the expected loss differences are unchanged when using a conditionally unbiased proxy \( \hat{Y}_t \). Here, we consider the implications of Theorem 1 for statistical tests of ECPA, and show that exact robustness leads to valid tests (in the sense that size is kept) whose power increases for more accurate proxies.

To that end, we outline the framework of conditional predictive ability testing pioneered by Giacomini and White (2006), which extends the classical predictive ability tests of Diebold and Mariano (1995). Recall that interest in ECPA tests centers on the null hypothesis

\[
H_0 : E_{t-1}[d(Y_t, x_{1t}, x_{2t})] \equiv 0 \quad \text{for all } t = 1, 2, \ldots
\]

To test this conditional moment condition in a finite sample of length \( n \), Giacomini and White (2006) proposed to test the implication of \( H_0 \) that \( E[h_{t-1}d(Y_t, x_{1t}, x_{2t})] = 0 \) for a \( \mathcal{F}_{t-1} \)-measurable and \( \mathbb{R}^r \)-valued test function \( h_{t-1} \). They do so using the Wald-type test statistic

\[
T_n = n\hat{Z}_t\hat{\Omega}^{-1}_n\hat{Z}_t,
\]

where \( \hat{Z}_t = h_{t-1}(\hat{Y}_t, x_{1t}, x_{2t}) \), \( \hat{\Omega}_n = 1/n \sum_{t=1}^n \hat{Z}_t \hat{Z}_t' \), and \( \hat{\Omega}_n \) is an invertible and consistent estimator of \( \var{\sqrt{n}\hat{Z}_t} \). Under weak regularity conditions, Giacomini and White (2006, theor. 1) showed that \( T_n \) is asymptotically \( \chi^2_2 \)-distributed under \( H_0 \).

However, when \( Y_t \) is not observed, \( T_n \) cannot be computed and we have to rely on the feasible test statistic

\[
\hat{T}_n = n\hat{Z}_t\hat{\Omega}^{-1}_n\hat{Z}_t,
\]

where \( \hat{Z}_t = h_{t-1}(\hat{Y}_t, x_{1t}, x_{2t}) \), \( \hat{\Omega}_n = 1/n \sum_{t=1}^n \hat{Z}_t \hat{Z}_t' \), and \( \hat{\Omega}_n \) is an invertible and consistent estimator of \( \var{\sqrt{n}\hat{Z}_t} \). Arguing as before, \( \hat{T}_n \) follows a \( \chi^2_2 \)-distribution asymptotically under the proxy hypothesis

\[
\hat{H}_0 : E_{t-1}[d(\hat{Y}_t, x_{1t}, x_{2t})] \equiv 0 \quad \text{for all } t = 1, 2, \ldots
\]

Hence, in the presence of measurement error, \( \hat{T}_n \) can validly test \( H_0 \) if \( E_{t-1}[d(Y_t, x_{1t}, x_{2t})] \equiv E_{t-1}[d(\hat{Y}_t, x_{1t}, x_{2t})] \), that is, if \( H_0 \) and \( \hat{H}_0 \) are equivalent, which is obviously implied by our exact robustness property.

Under exact robustness to measurement error, \( H_0 \) implies that \( E[d(Y_t, x_{1t}, x_{2t})] = 0 \) by the law of iterated expectations (LIE). Thus, when \( L(\cdot, \cdot) \) is of the form given in Theorem 1, EPA tests of the hypothesis \( E[d(Y_t, x_{1t}, x_{2t})] = 0 \) can also be validly carried out using a conditionally unbiased proxy \( \hat{Y}_t \) of \( Y_t \).

**Remark 6.** For HF proxies in finance, Li and Patton (2018) established that equality of the expected loss differences is not strictly necessary for the validity of equal predictive ability tests under measurement error. Instead, it suffices that the expected loss differences converge at rate \( o(\sqrt{n}) \), which they
call the “convergence-of-hypotheses” condition. However, verification of the latter often requires nontrivial conditional
conditions. For example, when comparing volatility forecasts with high-frequency proxies, the sample size of intraday returns
for computing volatility proxies must diverge faster than the number of out-of-sample volatility forecasts. In macroeconomic
applications, where the sampling frequency cannot be arbitrarily increased, the “convergence-of-hypotheses” condition is
not applicable. One conclusion of our Theorem 1 is that the convergence-of-hypotheses condition is not required (because it
holds trivially) for mean forecasts when conditionally unbiased proxies are used.

Next, we derive the limit of \( \hat{T}_n \) under the local alternative
\[
H_{a, loc} : E[h_{t-1}d(Y_t, x_{1t}, x_{2t})] = \frac{\delta}{\sqrt{n}} \quad \text{for all } t = 1, 2, \ldots ,
\]
where \( \delta \in \mathbb{R}^d \). The magnitude of the local alternative, \( \delta/\sqrt{n} \), converges to the null hypothetical value of zero as \( n \to \infty \), thus making it harder for our test to reject \( H_0 \) for increasing \( n \). Since \( E[Z_t] = \delta/\sqrt{n} \) under \( H_{a, loc} \), \( Z_t \) depends on \( n \). We reflect this in our notation by writing \( Z_t = Z_{n,t} = (Z_{n,1}^{(1)}, \ldots , Z_{n,d}^{(d)})' \). Note that for \( \delta = 0 \), \( H_{a, loc} \) is strictly speaking not an alternative as it reduces to the null. However, the method of proof for deriving the asymptotic limit of \( \hat{T}_n \) is the same for all \( \delta \in \mathbb{R}^d \). Hence, we leave \( \delta \) unrestricted in \( H_{a, loc} \). Values of \( \delta \) with larger norm correspond to local alternatives with larger magnitudes. Note that under fixed alternatives, ECPA tests are consistent, that is, power converges to one no matter which conditionally unbiased proxy is used. Thus, only by considering local alternatives are we able to derive analytical results on the power of tests that use different proxies.

If the underlying loss function is exactly robust, then it follows under \( H_{a, loc} \) by the LIE that
\[
\frac{\delta}{\sqrt{n}} = E[h_{t-1}d(Y_t, x_{1t}, x_{2t})] = E[h_{t-1}E(d(Y_t, x_{1t}, x_{2t}) \mid \mathcal{F}_{t-1})] = E[h_{t-1}d(\hat{Y}_t, x_{1t}, x_{2t})]
\]
for any conditionally unbiased proxy \( \hat{Y}_t \). The property in (4), implied by exact robustness, is essential for our local power results, because it allows to uncover \( E[h_{t-1}d(Y_t, x_{1t}, x_{2t})] = \delta/\sqrt{n} \) via the \( \hat{Z}_{n,t} = h_{t-1}d(\hat{Y}_t, x_{1t}, x_{2t}) \). One key benefit of exact robustness is that it allows to compare “how difficult” it is to assess the relative merits of two forecasts, when only an imperfect proxy \( \hat{Y}_t \) is available. Note that even though for different proxies \( \hat{Y}_t \), the deviations under (4) are equal, better proxies may still give less variable \( \hat{Z}_{n,t} \), such that departures from \( H_0 \) of the form in (4) may be detected more easily. To investigate this, we derive the local power of the test based on \( \hat{T}_n \).

To do so, we make the following assumptions, which are similar to those of Giacomini and White (2006). As only some proxy variable \( \hat{Y}_t \) is available, we specify these assumptions in terms of \( \hat{Z}_{n,t} = h_{t-1}d(\hat{Y}_t, x_{1t}, x_{2t}) \). Nonetheless, the specific (conditionally unbiased) choices \( \hat{Y}_t = Y_t \) and \( \hat{Y}_t = x_2 \) are also allowed. Let \( \hat{W}_t = (\hat{Y}_t, X_t)' \), where the \( \mathcal{F}_{t-1} \)-measurable \( X_t \) is \( \mathbb{R}^s \)-valued and contains all predictors that the forecasts \( x_{1t} \) and \( x_{2t} \) are based on. For instance, \( X_t \) may contain lagged \( \hat{Y}_t \).

D1: \( \{ (\hat{W}_t, \hat{Y}_t)' \} \) is \( \alpha \)-mixing of size \( -2r/(r-2) \) or \( \phi \)-mixing of size \( -r/(r-1) \).
D2: \( E[\hat{Z}_{n,t}^2(t)] \leq \Delta_Z < \infty \) for \( i = 1, \ldots , q \) and \( r \) from D1.
D3: \( \Omega_n = \text{var}(n^{-1/2} \sum_{t=1}^n \hat{Z}_{n,t}) \to \Omega, \) as \( n \to \infty \), where \( \Omega \) is positive definite.
D4: The forecasts \( x_{1t} = f_{n,t}^{(1)}(X_t, \ldots , X_{t-m_1+1}) \) and \( x_{2t} = f_{n,t}^{(2)}(X_t, \ldots , X_{t-m_2+1}) \) are measurable functions of a finite number of predictors, where \( m = \max(m_1, m_2) \leq m < \infty \).

Our conditions D1–D4 are standard in the literature, and are essentially those of Giacomini and White (2006, theor. 3). Assumption D4, that forecasts are based on a finite number of lags of the predictor variables, is a technical convenience used to expedite the proof of Theorem 2. Nonetheless, this assumption accommodates many different estimation schemes, where \( m_1 \) and \( m_2 \) may be deterministic or data-driven (and, hence, stochastic). We refer to Giacomini and White (2006) for more detail. Note that in D4 we suppress the dependence of the forecasts on \( n \) for notational brevity.

Theorem 2. Let Assumptions D1–D4 hold. Then, if \( \hat{\Omega}_n - \Omega_n \to op(1) \), it holds for any exactly robust loss function under \( H_{a, loc} \) that
\[
\hat{T}_n \overset{d}{\to} \chi^2_{q}(\delta \Omega^{-1} \delta), \quad \text{as } n \to \infty,
\]
where \( \chi^2_{q}(c) \) denotes a \( \chi^2_{q} \)-distribution with noncentrality parameter \( c \in \mathbb{R} \).

In the special case when \( Y_t = \hat{Y}_t \), Theorem 2 is a refinement of the consistency under fixed alternatives established by Giacomini and White 2006 (Theorem 2). For \( \delta = 0 \), Theorem 2 shows that \( \hat{T}_n \) has a standard \( \chi^2 \)-limit under \( H_0 \). Moreover, it implies that the asymptotic local power (ALP), that is, the asymptotic probability of rejecting \( H_0 \) under \( H_{a, loc} \), is
\[
P \{ \chi^2_{q}(\delta \Omega^{-1} \delta) > \chi^2_{q,1-\tau}(0), \}
\]
where \( \chi^2_{q,1-\tau}(0) \) is the \( (1-\tau) \)-quantile of the \( \chi^2_{q}(0) \)-distribution and \( \tau \in (0, 1) \) is the significance level of the test. This shows that the ALP depends on the magnitude of the local alternative (via \( \delta \)) and on the proxy via the limit of \( \Omega_n \) \( \Omega_n = \text{var}(n^{-1/2} \sum_{t=1}^n \hat{Z}_{n,t}) \). Thus, more precise proxies that give less variable \( \hat{Z}_{n,t} \) lead to tests with higher local power. We shed more light on this in the next subsection.

To make the test based on \( \hat{T}_n \) operational, we need a consistent estimator \( \hat{\Omega}_n \) of \( \Omega_n \). To that end, we consider a heteroscedasticity and autocorrelation consistent (HAC) estimator
\[
\hat{\Omega}_n = \frac{1}{n} \sum_{t=1}^n \hat{Z}_{n,t} \hat{Z}_{n,t}' + \frac{1}{n} \sum_{h=1}^{m_n} \sum_{t=h+1}^n \lambda \{ \hat{Z}_{n,t-h} \hat{Z}_{n,t-h}' \},
\]
where \( m_n \) is a sequence of integers, and \( w_{n,h} \) is a scalar triangular array of weights (Newey and West 1987). We restrict \( m_n \) and \( w_{n,h} \) as follows:
D5: The sequence of integers \( m_n \) satisfies \( m_n \to \infty \) and \( m_n = o(n^{1/4}) \), as \( n \to \infty \).
Proposition 1. Let Assumptions D1–D6 hold. Then, it holds under $H_{\text{a,loc}}$ that $\hat{\Omega}_n - \Omega_n = o_p(1)$, as $n \to \infty$.

The estimator $\hat{\Omega}$ is the omnibus choice. It works for EPA tests (where $H_0$ reduces to $E[d(Y_{t_1}, x_{1_1}, x_{2_2})] = 0$ by the LIE) but also for ECPA tests. For the latter tests, simpler estimators may be used, because—under $H_0$—the sequences $(h_{t-1}^{-1}(\hat{Y}_t, x_{1_1}, x_{2_2}), F_t)$ are margingale differences and thus, uncorrelated. This implies that $\Omega_n$ simplifies to $\Omega_n = 1/n \sum_{n=1}^t E[Z_{n,t-1}]$, rendering $\hat{\Omega}_n = 1/n \sum_{n=1}^t \hat{Z}_{n,t-1}$ the estimator of choice. However, even for EPA tests of one-step-ahead forecasts considered here, Diebold and Mariano (1995) recommended to use $m_n = 0$ in (6) (with an empty sum defined to be zero). Thus, using $\hat{\Omega}_n = 1/n \sum_{n=1}^t \hat{Z}_{n,t-1}$ for both EPA and ECPA tests seems reasonable, and we opt for this choice in the numerical experiments in Section 3.

2.3. Finding Optimal Proxies

Under the local alternative $H_{\text{a,loc}}$, Theorem 2 shows that the test’s asymptotic local power is maximized by choosing a proxy $\hat{Y}_t$ that minimizes $\Omega$. We discuss such choices in the following by using the linearity of the Bregman loss functions in $Y_t$. Specifically, we have

Proposition 2. Suppose that the assumptions of Theorem 1 hold and that any of (a)–(d) in that theorem are in force. Then, it holds that

$$\Omega_n = \frac{1}{n} \sum_{n=1}^t \text{var}(h_{t-1}^{-1}(x_{1_1}, x_{2_2}))$$

$$+ \frac{1}{n} \sum_{n=1}^t \text{cov}(h_{t-1}^{-1}(\hat{Y}_t), h_{t-1}^{-1}(x_{1_1}, x_{2_2})), h_{t-1}^{-1}(x_{1_1}, x_{2_2})),$$

where $h_{t-1} = d\phi(x_{1_1}) - d\phi(x_{2_2})$. Furthermore, if the covariance terms vanish asymptotically,

$$\Omega_n = \frac{1}{n} \sum_{n=1}^t \text{var}(h_{t-1}^{-1}(x_{1_1}, x_{2_2}))$$

$$+ \frac{1}{n} \sum_{n=1}^t \text{cov}(h_{t-1}^{-1}(\hat{Y}_t), h_{t-1}^{-1}(x_{1_1}, x_{2_2})) = o(1),$$

with an asymptotic lower bound of $\Omega^* = \lim_{n \to \infty} 1/n \sum_{n=1}^t \text{var}(h_{t-1}^{-1}(x_{1_1}, x_{2_2}))$ (in the sense that $\Omega - \Omega^*$ is positive semi-definite), which is attained if and only if $\hat{Y}_t = x_t^*$ a.s.

The covariance terms in (7) vanish (and, hence, (8)) holds if the $\{h_{t-1}^{-1}(Y_t, x_{1_1}, x_{2_2})\}$ are serially uncorrelated, which holds under $H_0$. But (8) may also hold under $H_{\text{a,loc}}$ as the derivation of $\Omega$ in Proposition 4 of Appendix C shows. If (8) holds, the generally infeasible choice of $\hat{Y}_t = x_t^*$ gives an upper bound for the local test power as then, $\text{var}_{t-1}(\hat{Y}_t) \equiv 0$, such that the second (positive semi-definite) term on the right-hand side of (8) vanishes. This is very intuitive: It is easiest to assess the relative merits of two forecasts $x_1$ and $x_2$ if they can be compared against the forecast target $x_t^*$ itself. Generally, this supports the intuition that it is easier to distinguish between two forecasts if the target variable is approximated more precisely. We stress once again that this result relies on the newly introduced exact robustness concept. While comparing forecasts with $x_t^*$ itself is best from a theoretical point of view, we are not aware of a practical situation where $x_t^*$ is observable, even ex post.

Since the test power decreases with increasing average variances $\text{var}_{t-1}(\hat{Y}_t)$ (in terms of the Loewner order), the proxy $\hat{Y}_t$ with smallest possible conditional variance should be employed in practice. However, these conditional variances are generally unknown in practice. In this case, one may resort to a proxy that is best from a theoretical point of view. We now discuss this exemplarily in the univariate case (k = 1) for the volatility forecasting example of Remark 2, and for our macroeconomic application of Section 4.

In the macroeconomic application, $Y_t$ denotes true GDP growth. As discussed in the Motivation, true GDP growth cannot be observed, even ex post. However, several different proxies $\hat{Y}_t$ are available. Suppose, as is plausible, that there is some additive measurement error, such that $\hat{Y}_t = Y_t + \hat{e}_t$; see, for example, Faust, Rogers, and Wright (2005) for such a modeling approach. Here, the estimation error $\hat{e}_t$ is independent of $F_{t-1}$, that is, $\hat{e}_t$ cannot be predicted from information available at time $t - 1$. Then, $\text{var}_{t-1}(\hat{Y}_t) = \text{var}_{t-1}(Y_t) + \text{var}(\hat{e}_t)$. Hence, choosing the most accurately estimated proxy $\hat{Y}_t$ (i.e., one with smallest possible $\text{var}(\hat{e}_t)$) gives ECPA tests higher (local) power; see also Section 4. Observing a proxy $\hat{Y}_t$ closer to $x_t^*$ than $Y_t$—in the sense that $\text{var}_{t-1}(\hat{Y}_t)$ is closer to $\text{var}_{t-1}(x_t^*) (= 0)$ than $\text{var}_{t-1}(Y_t)$—seems delusive in this application.

While in the previous example, true GDP growth as the target variable is the best proxy, one can sometimes get closer to the ideal $x_t^*$. To illustrate this, consider the classical situation of forecasting the conditional variance $x_t^* = \text{var}_{t-1}(r_t)$ of the univariate log-return $r_t$ on a risky asset with $E[r_t | F_{t-1}] = 0$. In this case, $x_t^* = E[r_t^2 | F_{t-1}]$, such that $Y_t = r_t^2$ is the natural target against which to compare forecasts. However, for a standard diffusion process, as in Andersen et al. (2003), one can show that $x_t^* = E[r_t^2 | IV_t]$ with $IV_t$ the latent integrated variance. Let $\hat{Y}_t = RV_t$ be the realized variance (RV), which estimates $IV_t$ from HF returns. Then, under Proposition 3 of Andersen et al. (2003) and given that $RV_t$ is an unbiased estimator of $IV_t$, it holds that $E[r_t | Y_t] = E[r_t | \hat{Y}_t] = x_t^*$. In this setting, it is already well-documented that $\hat{Y}_t = RV_t$ exhibits much smaller conditional variance than the squared return $Y_t = r_t^2$. Hence, $RV_t$ should be favored for testing ECPA of volatility forecasts as shown by our Theorem 2 and Proposition 2. Indeed, this is common practice in the literature when comparing volatility forecasts (Blair, Poon, and Taylor 2001; Hansen and Lunde 2005; Laurent, Rombouts, and Violante 2013). However, our Theorem 2 and Proposition 2 are the first theoretical results supporting this practice. (We mention that Andersen and Bollerslev (1998) provided early arguments for the use of HF proxies in the absolute evaluation of volatility forecasts via Mincer–Zarnowitz.
We first consider the squared error (SE) as a robust loss function, that is, \( L(Y, x) = (Y - x)^2 \), which arises for \( \phi(x) = x^2 \) and \( a(Y) = -Y^2 \) in (2). As the SE loss elicits the mean, we have \( x_t^n = E_{t-1}[Y_{t}] = \mu(1 - \phi) + \phi Y_{t-1} \). Hence, \( x_{t1} \) equals the optimal forecast confounded by some additive noise with variance \( \sigma^2_1 \). This implies that the larger \( \sigma^2_1 \), the worse the forecast \( x_{t1} \). On the other hand, \( x_{2t} \) is a biased forecast, yet with no additive noise. Note that when \( \sigma^2_1 > \mu^2(1 - \phi^2) \), \( x_{t1} \) is the preferred forecast; see (13).

Following Giacomini and White (2006), we choose the \( F_{1t} \)-predictable test function \( h_{t1} = (1, Y_{t-1})' \). To assess the magnitude of the local alternative, Appendix C shows that

\[
E[\hat{h}_{t1}d(Y_{t}, x_{t1}, x_{2t})] = \left( E[d(Y_{t}, x_{t1}, x_{2t})] \right) \left( \sigma^2_1 - \mu^2(1 - \phi^2) \right) = \left( \xi/\sqrt{n} \right). 
\]

Thus, we simulate under the local alternative

\[
H_{\text{a,loc}} : E[h_{t1}d(Y_{t}, x_{t1}, x_{2t})] = \delta/\sqrt{n},
\]

where \( \delta = (\xi, \mu \xi)' \). For \( \xi = 0 \), our results correspond to size. To compute the theoretical ALP in (5), we calculate \( \Omega \) from D3 in Proposition 4 of Appendix C.

We run our simulations on a grid of \( \xi \) values in the interval \([-4, 4]\), and compare the empirical rejection frequencies based on 10,000 simulation replications with the theoretical ALP at a significance level of 5%. We do so for \( \mu = 1 \), \( \phi = 0.2 \), and \( \sigma^2 = 1 \). To investigate the impact of noise in the proxy, we consider several values of \( \sigma^2 \). Specifically, we choose \( \sigma^2 \) such that \( \zeta = \var(Y_t)/\sigma^2 \in \{1/5, 1/2, 1, 2, 5, \infty\} \) with \( \zeta = \infty \) corresponding to the case of no noise, that is, \( \sigma^2 = 0 \). The ratio \( \zeta \) can be interpreted as the signal-to-noise ratio (SNR) of the noisy target; for \( \zeta = \infty \), we accurately observe the target, while for \( \zeta = 1 \), the signal \( Y_t \) is of the same magnitude as the noise \( \hat{\epsilon}_t \).

The six panels in Figure 1 correspond to the six different values of \( \xi \). In each panel, the empirical rejection frequencies are displayed as a function of \( \xi \) for sample sizes \( n \in \{50, 100, 500\} \). The theoretical ALP is displayed as the black reference line. Note that by (5) the ALP curve is symmetric around \( \delta = 0 \) and, hence, also around \( \xi = 0 \). We draw two conclusions from Figure 1.

First, as predicted by Theorem 2, the empirical rejection frequencies converge to the ALP curve as \( n \to \infty \). In particular, no matter how noisy the proxy, size is approximately accurate. As could be expected from a local power result, the rejection frequencies do not increase monotonically with the sample size as is typically the case for fixed alternatives. Here, while for larger \( n \) rejections for \( \xi > 0 \) are more frequent in Figure 1, for smaller \( n \) rejections for \( \xi < 0 \) occur more often. Second, as also expected from Theorem 2, more precise proxies (i.e., proxies with higher \( \zeta \)) lead to easier discrimination between forecasts in the sense of higher power; see in particular the upper three panels. On the other hand, when \( \sigma^2_1 \to \infty \), the signal in \( Y_t \) is eventually buried by the noise in \( \hat{\epsilon}_t \), giving only close to trivial power in the lower three panels.

To summarize, these simulations show that for noisy target variables, ECPA tests are valid when using robust loss functions.
in the sense of unaffected rejection rates under the null hypothesis. However, even for robust loss functions, noisy targets negatively influence the test power, which suggests that applied researchers should use the most accurate proxy in forecast comparisons.

### 3.2. Illustration for Nonrobust Loss Functions

Now, we illustrate the key negative implication of Theorem 1, namely that if there is measurement error in the target (and, hence, only a conditionally unbiased proxy is available for comparing forecasts), then considering any functional different from the mean leads to nonrobust comparisons. We stress once again that we view this implication as the most relevant for practitioners, because when there is measurement error, it narrows down the scope of robust forecast comparisons to the mean functional.

Specifically, we consider $\alpha$-quantile forecasts (instead of mean forecasts) for $Y_t$ for $\alpha \in (0, 0.5)$. We again take $Y_t$ and $\tilde{Y}_t$ from (9) and (10). The forecasts are

$$x_{1t} = \mu(1 - \varphi) + \psi Y_{t-1} + \sigma_1 \Phi^{-1}(\alpha),$$

$$x_{2t} = \mu(1 - \varphi) + \psi Y_{t-1} + \sqrt{\sigma_1^2 + \sigma_2^2} \Phi^{-1}(\alpha),$$

where $\Phi(\cdot)$ denotes the standard normal distribution function. Therefore, $x_{1t}$ is the $\alpha$-quantile of $Y_t | F_t \sim N(\mu(1 - \varphi) + \psi Y_{t-1}, \sigma_1^2)$, and $x_{2t}$ is the $\alpha$-quantile of $\tilde{Y}_t | F_t \sim N(\mu(1 - \varphi) + \psi Y_{t-1}, \sigma_1^2 + \sigma_2^2)$. To compare the $\alpha$-quantile forecasts, we use the standard tick loss function $L(Y, x) = (1(Y \leq x) - \alpha)(x - Y)$; see Gneiting (2011). Straightforward computations exploiting Example 2.14 in McNeil, Frey, and Embrechts (2015) yield

$$E[d(Y_t, x_{1t}, x_{2t})] = \sigma_1 \left\{ \Phi'(\Phi^{-1}(1 - \alpha)) - \Phi'(\Phi^{-1}(1 - \tilde{\alpha})) \right\} + \Phi^{-1}(1 - \tilde{\alpha})|\alpha - \tilde{\alpha}|,$$

$$E[d(\tilde{Y}_t, x_{1t}, x_{2t})] = \sqrt{\sigma_1^2 + \sigma_2^2} \left\{ \Phi'(\Phi^{-1}(1 - \tilde{\alpha})) - \Phi'(\Phi^{-1}(1 - \alpha)) \right\} + \Phi^{-1}(1 - \tilde{\alpha})|\alpha - \tilde{\alpha}|,$$

where $\tilde{\alpha} \in (0, 1)$ satisfies $\sqrt{\sigma_1^2 + \sigma_2^2} \Phi^{-1}(\alpha) = \sigma_1 \Phi^{-1}(\tilde{\alpha})$, $\tilde{\alpha} \in (0, 1)$ satisfies $\sqrt{\sigma_1^2 + \sigma_2^2} \Phi^{-1}(\alpha) = \sigma_1 \Phi^{-1}(\alpha)$, and $\Phi'(\cdot)$ is the standard normal density.

Figure 2 plots both $E[d(Y_t, x_{1t}, x_{2t})]$ and $E[d(\tilde{Y}_t, x_{1t}, x_{2t})]$ for different values of $\alpha$. For ordering robust comparisons, we expect identical signs of both quantities, and for exactly robust comparisons, the two quantities should even be equal. Yet, Figure 2 shows that the expected loss differences have opposite signs for all $\alpha \in (0, 1/2)$, illustrating that comparisons of quantile forecasts are sensitive to measurement error. Notice that the forecasts $x_{1t}$ and $x_{2t}$ coincide for the special case of $\alpha = 1/2$ as $\Phi^{-1}(1/2) = 0$.

The fact that there may be sign changes for the expected loss differences under measurement error implies that, in sufficiently large samples, inference based on $d(Y_t, x_{1t}, x_{2t})$ may lead to the erroneous conclusion that, say, $x_{2t}$ is superior to $x_{1t}$ when in fact the opposite is true. This casts some doubt on the results of quantile forecast comparisons for GDP growth, as carried
out by Brownlees and Souza (2021). With the rising popularity of the GaR, which is the conditional \( \alpha \)-quantile of GDP growth, such comparisons will likely become more relevant in the future (Adrian, Boyarchenko, and Giannone 2019; Adrian et al. 2021). Our Theorem 1 suggests that such comparisons need to be interpreted with caution due to the nonrobustness of the quantile functional. The only way to alleviate this concern is by using a proxy that resembles the actual target variable as much as possible. We investigate this and further finite-sample implications of the nonrobustness of quantiles in Appendix D, where we consider the specific case of the median (i.e., \( \alpha = 1/2 \)).

4. Comparing Forecasts for U.S. GDP Growth

4.1. Data

As a measure of total real activity, GDP is arguably the most important macroeconomic indicator. However, as pointed out in the Motivation, GDP—and, by extension, GDP growth—can only be measured with error. In this application, we focus on continuously compounded growth rates for U.S. GDP, denoted by \( \Delta GDP_t \). To illustrate Theorem 1, we demonstrate the difficulty of robustness when evaluating mean and quantile forecasts in Section 4.2. To illustrate Theorem 2, Section 4.3 provides evidence for the increasing test power of mean forecast comparisons when using more exact proxies. All datasets used in this application are downloaded from https://www.philadelphiafed.org/surveys-and-data/real-time-data-research.

To describe the growth proxies \( \hat{Y}_t \), recall that there are three ways to compute GDP: the production, income, and expenditures approach; see Landefeld, Seskin, and Fraumeni (2008, tab. 1). All these methods may be regarded as providing estimates of the true latent value of GDP. We use proxies based on the income or expenditures approach, or a combination of the two.

Our first \( \Delta GDP \) proxy is based on expenditure-side estimates and the second one on the income approach. We denote them by \( \Delta GDP_E \) and \( \Delta GDP_I \). While proxies based on the income method feature less prominently in economics, Nalewai (2010) nonetheless found \( \Delta GDP_I \) to better reflect the growth in real economic activity during the business cycle. As our third proxy, we use the “GDP Plus” approach of Aruoba et al. (2016), which combines estimates from the income and expenditure side. The authors argue that \( \Delta GDP_I \) and \( \Delta GDP_E \) provide complementary information on GDP growth. The recent popularity of \( \Delta GDP_+ \) is documented by the Federal Reserve Bank of Philadelphia reporting it alongside the more classical measures \( \Delta GDP_E \) and \( \Delta GDP_I \). The methodology behind \( \Delta GDP_+ \) has also spurred the development of new GDP proxies (Jacobs et al. 2022).

In more detail, Aruoba et al. (2016) proposed a dynamic factor model, where \( \Delta GDP_E \) and \( \Delta GDP_I \) both load on the single (latent) factor \( \Delta GDP \)

\[
\left( \begin{array}{cc} 
\Delta GDP_{E,t} \\
\Delta GDP_{I,t} 
\end{array} \right) = \left( \begin{array}{cc}
1 \\
1
\end{array} \right) \Delta GDP_t + \left( \begin{array}{c} 
\varepsilon_{E,t} \\
\varepsilon_{I,t}
\end{array} \right),
\]

where \( (\varepsilon_{E,t}, \varepsilon_{I,t})' \) are assumed to be iid with mean zero, implying in particular conditional unbiasedness of the growth proxies, that is, \( E_{t-1}[\Delta GDP_{E,t}] = E_{t-1}[\Delta GDP_{I,t}] = E_{t-1}[\Delta GDP_t] \). The authors extract GDP growth estimates, denoted \( \Delta GDP_+ \), by using the Kalman smoother. Crucially for us, Aruoba et al. (2016) argued that the Kalman filter extractions \( \Delta GDP_+ \) can be regarded as more precise approximations of true growth than either \( \Delta GDP_E \) or \( \Delta GDP_I \).

We consider the three GDP proxies from 1985Q1 until 2019Q3 to avoid structural breaks in our sample due to the Great Moderation and the Corona crisis. Aruoba et al. (2016) noted in their conclusion that early vintages of \( \Delta GDP_I \)—and hence of \( \Delta GDP_+ \)—often provide less accurate information than their counterparts from the expenditure side. Therefore, we use the latest data revision (also called the latest “vintage”) as of September 30, 2020, for our proxies.

One may wonder whether the differences in our proxies are large enough during the sampling period to suspect substantially altered results of ECPA tests. To shed light on this, Figure 3 displays the three GDP proxies over time. We see a clear joint behavior while the exact measurements differ—sometimes even substantially. For instance, in 2007Q1 the expenditure approach suggests a growth rate of roughly 2% whereas the income and GDP Plus approaches indicate a declining economy with rates of \(-3.0% \) and \(-1.3% \) respectively. Hence, the results of ECPA tests may plausibly depend on which proxy is used in the comparison.
To illustrate Theorems 1 and 2 in Sections 4.2 and 4.3, we require the proxies to be conditionally unbiased, and Appendix E.1 provides solid evidence for this.

4.2. (Non)Robust Evaluation of Mean and Quantile Forecasts

Here, we illustrate the implication of Theorem 1 that an exactly robust forecast evaluation via strictly consistent loss functions is possible if and only if the target is the mean functional. We do so by comparing the robustness of loss functions for mean and quantile forecasts when evaluated by different GDP proxies. While traditional forecasts for GDP growth target its conditional mean, quantile forecasts gained a lot of attention recently in the emerging GaR literature (Adrian, Boyarchenko, and Giannone 2019; Adrian et al. 2021; Brownlees and Souza 2021). The GaR is simply a conditional \( \alpha \)-quantile of GDP growth rates, where \( \alpha \) is close to zero, for example, \( \alpha \in [0.05, 0.1] \).

We consider mean forecasts generated by an autoregressive (AR) model of order one (\( x_{t1} \) in our notation) against historical forecasts (\( x_{t2} \) in our notation). The latter are simply the mean over the training window. We use similar models to generate quantile forecasts. First, we use an autoregressive quantile model of order one (Koenker and Xiao 2006), that is, we estimate a quantile regression at level \( \alpha \) of GDP growth on a constant and its lagged value. Second, we use the sample quantiles over the estimation window. These forecasts are often used as benchmark models for mean and quantile forecasts of GDP growth; see, for example, Stock and Watson (2003), Adrian, Boyarchenko, and Giannone (2019), and Brownlees and Souza (2021). We use a training window size of 40 quarters, and in order to mimic real-time forecasting, we use data from the first vintage of the expenditure side GDP estimates for model estimation. This choice also avoids giving an unrealistic advantage to any of the proxies used for the evaluation. Additionally, reasonably precise expenditure-side estimates are available earlier than estimates from the income-side.

Table 1 reports loss differences and \( p \)-values of the respective Diebold and Mariano (1995) tests of EPA. The mean forecasts are compared by the robust SE loss and the quantile forecasts by the nonrobust tick loss function. Both choices are strictly consistent for the respective functional and are standard choices in the literature. By Theorem 1 and the robustness of the SE loss, we expect consistent comparisons for mean forecasts even for different (conditionally unbiased) proxies. In contrast, differing evaluation results for quantile forecasts can arise for the nonrobust tick loss.

The results in Table 1 confirm these expectations: All three proxies yield congruent results for mean forecasts, whereas they sometimes differ in their evaluation of quantile forecasts. For \( \alpha = 0.05 \), the loss differences are all negative and comparable. This shows that the nonrobustness of the tick loss only implies that incoherent forecast rankings can occur in some instances, but they do not necessarily have to occur in all comparisons. For \( \alpha = 0.1 \), however, the conclusions differ for different proxies. The expenditure approach favors the historical model weakly, whereas the other two proxies favor the quantile autoregression significantly. The latter fact may be explained by the observation of Aruoba et al. (2016) that \( \Delta GDP_e \), loads more heavily on \( \Delta GDP_p \) than on \( \Delta GDP_e \). Thus, more similar results are expected for \( \Delta GDP_e \) and \( \Delta GDP_p \).

Our results demonstrate that one should exercise caution in interpreting quantile forecast comparisons in the presence of measurement error, such as in the GaR forecast comparison of Brownlees and Souza (2021). In particular, the results should be interpreted only with respect to the employed proxy variable and are not necessarily valid for other proxies or the true, latent target variable.

4.3. Power Comparison

We now restrict attention to mean forecasts such that the power results of Sections 2.2 and 2.3 apply. We evaluate forecasts using the same proxies as before and we investigate if—as indicated by Theorem 2 and Proposition 2—it is indeed easier to discriminate between two forecasts when a more precise proxy is used.
First, we use the autoregressive mean forecasts from the previous subsection, however, with forecast horizons of one, two and four quarters. These forecasts are directed at the mean functional by construction. The second set of growth forecasts is taken from the SPF. Since 1968, the SPF publishes quarterly forecasts—prepared by private sector economists—of macroeconomic variables in the United States. The SPF is widely used for forecast comparisons in the academic literature (e.g., Campbell 2007; Engelberg, Manski, and Williams 2009). Specifically, we consider the (cross-respondent) mean GDP growth predictions of the SPF. Appendix E.2 shows that the SPF forecasts can be interpreted as rational mean forecasts by using the method of Dimitriadis, Patton, and Schmidt (2021b).

Our goal is to show that ECPA tests can more easily discriminate between two forecasts $x_{11}$ and $x_{21}$ if a more precise proxy is used. To this end, we require one forecast to be clearly superior to the other. Hence, we employ forecasts for the same quarter $t$ (from either the AR(1) model or the SPF), however with varying horizons, that is, forecasts for $\Delta GDP_t$ issued at different time points $t - \tau$. Holzmann and Eulert (2014) formally showed that forecasts based on larger information sets are superior to those based on smaller information sets. The forecast with the shorter horizon ($x_{11}$) naturally nests the information set of the longer horizon forecast ($x_{21}$), implying superiority of the former. As we consider multi-step ahead forecasts, we employ a HAC covariance estimator (Newey and West 1987).

Table 2 shows $p$-values of ECPA tests (based on $\hat{T}_0$) for the three combinations of one-, two-, and four-quarter ahead GDP growth forecast. It does so individually for the AR(1) and the SPF forecasts in Panels A and B. These forecasts are evaluated by the proxies $\Delta GDP_E$, $\Delta GDP_P$, and $\Delta GDP_P^+$, respectively. As test functions $h_{t-1}$, we use a constant only (Inst. 1), and a constant jointly with the loss difference of the forecast, lagged by the horizon of the shorter of the two forecast horizons (Inst. 2).

We find that all loss differences in Table 2 are negative, implying—as expected—better predictive ability of the shorter horizon forecasts. The similarity of the average loss differences throughout the proxies again illustrates the exact robustness of the SE loss. For both panels, we find substantially lower $p$-values for the “GDP Plus” approach, which is claimed to be superior in the literature. For example, “GDP Plus” exhibits the smallest $p$-value in 11 out of the 12 cases, where the increase in significance upon the expenditure approach is particularly pronounced. We explain this by the theoretically higher test power for the more accurately estimated proxy variable. The fact that the results are more similar for $\Delta GDP_E$ and $\Delta GDP_P^+$ than for $\Delta GDP_E$ and $\Delta GDP_P$ may once again be explained by the fact that $\Delta GDP_P^+$ loads more heavily on $\Delta GDP_P$ (Aruoba et al. 2016).

5. Conclusion

In this article, we answer the following question: what can we learn from forecast comparisons if the target variable cannot be observed precisely, and only mismeasured proxy variables are available? We show that the classical forecast evaluation tools of loss functions and inference thereon can be used without modification if (a) a conditionally unbiased proxy is available and (b) the target functional is the conditional mean. In contrast, for other target functionals such as, for example, the conditional median, the use of approximated proxy variables generally distorts the loss differences and hence, inference in tests for ECPA.

This leads to the perhaps surprising conclusion that when evaluating forecasts in the presence of measurement error, the mean is “more robust” than the median, whereas the converse is well-known in classical estimation theory. Hence, our work should encourage practitioners to use the mean as target functional in forecasting settings that are prone to measurement error, such as in our empirical application on GDP growth. Further applications are widespread and include, among others, macroeconomic variables such as inflation rates, volatility forecasts in finance, meteorological quantities such as precipitation or wind speeds, and case or death counts in infectious disease forecasting.

We further demonstrate that even though standard inference in classical ECPA tests for mean forecasts is valid under measurement error, the test’s (local) power decreases with the magnitude of the error. This gives theoretical content to the empirical observation that “[a]lthough consistency of the ordering is ensured by an appropriate choice of the loss function independently of the quality of the proxy, a high precision proxy allows to efficiently discriminate between models” (Laurent, Rombouts, and Violante 2013, p. 7). We confirm this in Monte Carlo experiments and provide an empirical illustration on GDP growth. We emphasize that this power increase afforded by more precise proxies is particularly important in economics, where sample sizes are often limited due to low-frequency data collection, structural breaks, etc.

Table 2. Loss differences and $p$-Values of ECPA tests for multiple GDP measurements.

| Proxy | 1Q vs. 2Q Ahead | 1Q vs. 4Q Ahead | 2Q vs. 4Q Ahead |
|-------|----------------|----------------|----------------|
|       | $p$-value | $p$-value | $p$-value |
|       | Diff. Inst. 1 Inst. 2 | Diff. Inst. 1 Inst. 2 | Diff. Inst. 1 Inst. 2 |
| Panel A: Autoregressive Mean Forecasts | | | |
| $\Delta GDP_E$ | $-0.635$ | $0.219$ | $0.385$ | $-0.936$ | $0.178$ | $0.286$ | $-0.270$ | $0.185$ | $0.367$ |
| $\Delta GDP_P$ | $-0.980$ | $0.079$ | $0.065$ | $-1.298$ | $0.042$ | $0.005$ | $-0.250$ | $0.255$ | $0.254$ |
| $\Delta GDP_P^+$ | $-0.891$ | $0.069$ | $0.056$ | $-1.217$ | $0.040$ | $0.003$ | $-0.273$ | $0.127$ | $0.225$ |
| Panel B: SPF Forecasts | | | |
| $\Delta GDP_E$ | $-0.509$ | $0.118$ | $0.251$ | $-1.286$ | $0.094$ | $0.129$ | $-0.777$ | $0.091$ | $0.151$ |
| $\Delta GDP_P$ | $-0.557$ | $0.117$ | $0.245$ | $-1.304$ | $0.092$ | $0.096$ | $-0.747$ | $0.093$ | $0.078$ |
| $\Delta GDP_P^+$ | $-0.531$ | $0.093$ | $0.182$ | $-1.258$ | $0.086$ | $0.063$ | $-0.727$ | $0.086$ | $0.129$ |

NOTE: This table reports the loss differences (column “Loss Diff.”) together with the $p$-values of the tests of ECPA based on the two instrument choices given in the text for the different comparisons given in Panels A and B. The first column panel reports results for the comparison of one- against two-quarter ahead forecasts, whereas the second and third column panels compare one- against four-, and two- against four-quarter ahead forecasts.
Supplementary Materials

In the Supplementary Material, Appendix A establishes a nonrobustness result for quantile forecasts under any “resemblance condition” on the proxy. Appendices B and C contain all proofs and some further technical derivations. Appendix D provides additional simulation results for non-robust loss functions and Appendix E gives additional details for the empirical application. The supplementary material also contains R code to reproduce the numerical results in this article.

Acknowledgments

The authors are indebted to two referees and the associate editor, whose suggestions led to significant improvements of the article. Furthermore, the authors would like to thank Tobias Fissler, Christoph Hanck, Onno Kleen and Andrew Patton for their very insightful comments on an earlier version of this article. The authors are also grateful to seminar participants at the University of Bonn and the Heidelberg Institute for Theoretical Studies (HITS), in particular Tilmann Gneiting and Johanna Ziegel. Thanks are also due to the HITS for financial support. Moreover, the first author thankfully acknowledges support of the German Research Foundation (DFG) through project HO 6305/1-1.

References

Abowd, J. M., and Stinson, M. H. (2013), “Estimating Measurement Error in Annual Job Earnings: A Comparison of Survey and Administrative Data,” *The Review of Economics and Statistics*, 95, 1451–1467. [364]

Adrian, T., Boyarchenko, N., and Giannone, D. (2019), “Vulnerable Structure of Growth-at-Risk,” *Econometrica*, vol. 87, 1733–1765. [365, 373, 374]

Adrian, T., Grinberg, F., Liang, N., Malik, S., and Yu, J. (2021), “The Term Structure of Growth-at-Risk,” *American Economic Journal: Macroeconomics*, available at https://www.aeaweb.org/articles/pdf/doi/10.1257/ mac.20180428. [366, 373, 374]

Andersen, T. G., and Bollerslev, T. (1998), “Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts,” *International Economic Review*, 39, 885–905. [367, 370]

Andersen, T. G., Bollerslev, T., Christoffersen, P. F., Diebold, F. X. (2013), “Financial Risk Measurement for Financial Risk Management,” in *Handbook of the Economics of Finance* (Vol. 2, 1st ed.), eds. G. M. Constantinides, M. Harris, R. M. Stulz, Amsterdam: Elsevier, pp. 1127–1220. [364, 367]

Andersen, T. G., Bollerslev, T., Diebold, F. X., Laby, P. (2003), “Modeling and Forecasting Realized Volatility,” *Econometrica*, 71, 579–625. [370, 371]

Aruoba, S. B., Diebold, F. X., Nalewaik, J., Schorfheide, F., and Song, D. (2016), “Improving GDP Measurement: A Measurement-Error Perspective,” *Journal of Econometrics*, 191, 384–397. [364, 365, 373, 374, 375]

Banerjee, A., Guo, X., and Wang, H. (2005), “On the Optimality of Conditional Expectation as a Bregman Predictor,” *IEEE Transactions on Information Theory*, 51, 2664–2669. [364, 365]

Blair, B. J., Poon, S. H., and Taylor, S. J. (2001), “Forecasting S&P 100 Volatility: The Incremental Information Content of Implied Volatilities and High-Frequency Index Returns,” *Journal of Econometrics*, 105, 5–26. [370]

Bräuning, F., and Koopman, S. J. (2014), “Forecasting Macroeconomic Variables Using Collapsed Dynamic Factor Analysis,” *International Journal of Forecasting*, 30, 572–584. [365, 368]

Brownlee, C., Souza, A. B. M. (2021), “Backtesting Global Growth-at-Risk,” *Journal of Economics* 116, 112–130. [365, 373, 374]

Campbell S. D. (2007), “Macroeconomic Volatility, Predictability, and Uncertainty in the Great Moderation: Evidence From the Survey of Professional Forecasters,” *Journal of Business & Economic Statistics*, 25, 191–200. [375]

Clark, T., and McCracken, M. (2001), “Tests of Equal Forecast Accuracy and Encompassing for Nested Models,” *Journal of Econometrics*, 105, 85–110. [364]

Clark, T. E., and McCracken, M. W. (2009), “Tests of Equal Predictive Ability With Real-Time Data,” *Journal of Business & Economic Statistics*, 27, 441–454. [364]