Survey expectations and adjustments for multiple testing

Clements, M. P. ORCID: https://orcid.org/0000-0001-6329-1341 (2024) Survey expectations and adjustments for multiple testing. Journal of Economic Behavior and Organization, 224. pp. 338-354. ISSN 2328-7616 doi: 10.1016/j.jebo.2024.06.009 Available at https://centaur.reading.ac.uk/116781/

Journal of Economic Behavior and Organization 224 (2024) 338–354

Contents lists available at ScienceDirect

Journal of Economic Behavior and Organization

journal homepage: www.elsevier.com/locate/jebo

Research paper

Survey expectations and adjustments for multiple testing

Michael P. Clements

ICMA Centre, Henley Business School, University of Reading, United Kingdom

ARTICLE INFO

JEL classification:
C12
C53
E37

Keywords:
Multiple tests
Survey expectations
Family-wise error rates
False discovery rates

ABSTRACT

Testing hypotheses regarding how individual survey respondents form their expectations is susceptible to the multiple testing problem. The probability of falsely rejecting the null hypothesis for one or more respondents will exceed the nominal single-hypothesis significance level. The Bonferroni correction and related approaches control the family-wise error rate, but are conservative and result in low power when the null hypotheses are false.

We compare controlling the family-wise error rate with the effect of controlling the false discovery rate and the false discovery proportion, in terms of the conclusions we draw about forecaster behaviour.

The effects of adjustments for multiple testing are investigated for tests of weak efficiency and the over-reaction hypothesis, for beliefs about the persistence of shocks to output growth, and for the accuracy of survey respondents' perceptions of the uncertainty they face.

1. Introduction

Studies of the properties of the expectations of survey respondents at the individual level involve multiple hypothesis tests. In this paper we consider the effects of adjustments or corrections for multiple testing (MT) on findings in the literature on how professional forecasters form their expectations. The studies we consider test a null hypothesis for each of a number of survey respondents individually. The multiple testing problem arises because, as we increase the number of respondents being tested, the chances of rejecting a true null hypothesis will increase, simply because of chance. The classical approach to the MT problem is to control the family-wise error rate (FWER). The FWER is the probability of rejecting at least one true null hypothesis. Controlling the FWER requires carrying out the multiple tests in such a way that the probability of rejecting at least one true null hypothesis (the FWER) is less than or equal to a given probability, often take to be the significance level α at which each individual hypothesis test is carried out. The best-known approach to controlling FWER is the Bonferroni correction of the individual p-values, but this is known to be conservative, and alternatives to Bonferroni have been developed.

In the context of testing a given null hypothesis for each of a number of survey respondents, the classical approach of controlling the FWER may result in low power. Rather than controlling the FWER, control of the false discovery rate (FDR) - the *expected* false discovery proportion (FDP) (i.e., the proportion of rejections of the null which are false) — has been used in genomewide association studies (GWAS), and also in finance.[1] FDR control has been proposed for situations where only a small proportion of the

E-mail address: m.p.clements@reading.ac.uk.

[1] GWAS test the effects of thousands of genes — whether a particular gene being 'on' (or 'expressed') is associated with a particular disease. There is a null hypotheses for each gene, that disease status is independent of that gene. Even if the null of no effect of gene i were true for each gene, $1 = 1, \ldots, s$ (where the null for gene i, H_i, is no difference between the gene expression levels for the patients and controls for gene i), testing each null at the α percent level would result in a probability that at least one null is rejected of $1 - (1 - \alpha)^s$. This probability would be close to one for s of the magnitude typical in GWAS, but in excess of a half even for modest numbers of tests, such as $s = 20$, for the conventional $\alpha = 0.05$ significance level.

In empirical finance, applications that consider fund manager performance, or trading strategies, also result in very many null hypotheses: Chordia et al. (2020) consider over two million strategies (see also Barras et al., 2010, *inter alia*.) On the other hand, FDR has been applied to cases where s is much smaller

https://doi.org/10.1016/j.jebo.2024.06.009
Received 7 December 2023; Received in revised form 5 June 2024; Accepted 8 June 2024

null hypotheses are likely to be false — as in GWAS. GWAS hypotheses are typically 'disinterred in explorations', as opposed to being 'predesignated hypotheses'[2] with theoretical foundations. When most null hypotheses in the set are true, (unadjusted) inference will result in rejections of the null which are mainly false, resulting from 'luck', yielding a high proportion of false discoveries. Controlling the false discovery rate (the FDR) as in Benjamini and Hochberg (1995) controls the *expected* proportion of the hypotheses that are falsely rejected. That is, FDR controls the expected error rate for hypotheses for which the null is rejected. If the FDR is set at 5%, then no more than 5% of the rejected nulls would be expected to be true (that is, false discoveries). The FDR focuses on whether the rejections are legitimate, whereas FWER is concerned with the probability of obtaining false positives. Benjamini and Yekutieli (2001) (p. 1169) argue that 'The control of FDR assumes that when many of the tested hypotheses are rejected it may be preferable to control the proportion of errors rather than the probability of making even one error'.

A hallmark of survey-based macroeconomic expectations is that there appears to be persistent differences between individual respondents in a number of dimensions: for example, Jain (2019) finds heterogeneity in individuals' perceptions of the persistence in inflation, and Clements (2024) considers the heterogeneity of inflation expectations through the lens of individual Phillips curve models. Both studies consider the extent to which the heterogeneity depends upon the times at which the respondents were active participants. That is, of interest is understanding why such heterogeneity occurs: why we might reject a particular hypothesis for one individual, but not for another. Relatedly, there is an interest in the implications of whether a hypothesis holds, in terms of whether this is related to other behavioural hypotheses. For example, Clements (2022) considers whether the rejection of forecast efficiency suggests such individuals are more or less likely to make accurate forecasts, or to produce "contrarian" forecasts, than "efficient" forecasters. This means we are interested in identifying the individual forecasters for whom we reject, rather than in simply determining the number or proportion of forecasters for whom we reject.[3]

In this paper we are interested in the MT problem applied to survey expectations data, and so we do not consider the reasons behind the rejections.[4] We report the results as the proportion of rejections, simply as a way of summarizing and comparing the effects of different forms of MT correction.[5] Our interest is in obtaining a reliable estimate of the number or proportion of respondents whose behaviour is consistent with a particular null hypothesis. But it matters who those respondents are, and this motivates our testing approach. Given the low power of approaches which control FWER, we investigate whether less conservative approaches such as FDR provide more reliable estimates of the number of true and false nulls.

A key question is then whether making an allowance for MT (either using a type of FWER or FDR control) affects the conclusions we draw regarding various aspects of macroeconomic expectations formation. Does it make a material difference to our overall conclusions about whether survey respondents' forecasts are consistent with various postulates (e.g., respondents forecasts are weakly efficient, in the sense of Mincer and Zarnowitz (1969), or that they over-react to news)? Looking ahead to our findings, we find it does matter. For example, we reject the null hypothesis of forecast efficiency of short-horizon forecasts for around half of the respondents when no correction is made for MT. The standard method of controlling the FWER reduces the proportion of rejections from 1 in 2 to around 1 in 6. The making of inefficient forecasts becomes an affliction of a small minority of respondents, when we correct for MT in the usual way. Other approaches to correcting for MT have broadly similar effects. However, MT correction matters much less for long-horizon forecast efficiency. We contend that these findings, and those for the analysis of individual forecasters' perceptions of the persistence of output growth, and for an analysis of the accuracy of forecasters' perceptions of the uncertainty they face, could not have been foreseen in advance of carrying out the analysis. We are not aware of any applications of MT corrections to hypotheses concerning individual-level survey expectations, despite the widespread use of controls for MT in many disciplines and subject areas. Multiple hypothesis testing and the control of error rates is a potential issue in all disciplines that draw inferential conclusions from data.

We consider a number of ways of controlling FWER and FDR which appear most promising for multiple testing of individual survey respondents. We argue for the use of simple methods which can be applied without bootstrapping the underlying forecast data, and for the use of approaches which perform well when the test statistics (p-values) are dependent. We avoid simulation techniques such as bootstrapping, because of the complications that arise with missing data. In the macro survey data we use there are many missing forecasts, which would complicate the application of block-bootstrap methods (such as those advocated by Romano and Wolf, 2005). In our context p-values are likely to be dependent: we often consider forecast errors, which are based on a realization of the actual value common to all; individuals' forecasts draw on public information, etc.

The plan of the remainder of the paper is as follows. In Section 2 we briefly review FWER and FDR control. Section 3 presents a Monte Carlo study that examines the performance of the MT strategies for numbers of forecasters and forecast sample sizes typical

— see e.g., the re-analysis by Glickman et al. (2014) of two studies. In one $s = 28$, and in the other $s = 55$. Hence even though the number of hypotheses is typically far fewer for survey expectations than for GWAS, and in some finance applications, MT issues may need to be addressed.

[2] These terms are due to Mayo (2018), p. 275.

[3] We are grateful to a referee for making the point that if one were only interested in the proportion or number of rejections of a hypothesis, other approaches that sidestep the need to consider multiple hypotheses, and hence MT issues, might be worth considering. We take the view that generally one would be interested in *why* we reject for one individual rather than another, and whether that is influenced by the macro-environment at the time the individual was active, for example.

[4] However, as an example, the empirical application in Section 4.1 is taken from Clements (2022), who considers the implications of the rejection of efficiency for accuracy and disagreement.

[5] If we were to consider tests based on aggregate quantities, such as regressions of mean forecast errors on revisions to mean forecasts, as in Coibion and Gorodnichenko (2012, 2015), then MT issues would not arise, but we would not necessarily be testing the behavioural hypotheses of interest (as discussed in Section 4.1).

of those available in surveys of macroeconomic expectations. The data generating process and models in the Monte Carlo are based on the empirical application of Section 4.1 illustration. Section 4.1 applies multiple-testing adjustments to individual-level tests of forecast efficiency, and to tests of the over-reaction hypothesis. Section 4.2 applies multiple-testing adjustments to individual-level tests of whether professional forecasters believe output shocks are permanent. Section 4.3 applies multiple-testing adjustments to tests of the equality of *ex ante* and *ex post* uncertainty at the individual level. These applications are of individual-level testing of hypotheses concerning expectations formation by professional forecasters.[6] Section 5 concludes and summarizes our findings on the effects of multiple-testing adjustments for macro survey expectations.

2. Controlling error rates

We provide a brief review of methods of controlling error rates. These methods can all be applied relatively simply, in that they do not require bootstrapping. They control either the FWER or the FDR, and do so allowing for different types of dependence in the p-values.

2.1. FWER control

Given a set of null hypotheses, H_i, $i = 1, \dots, s$, let p_i, $i = 1, \dots, s$, denote the p-values associated with these hypotheses. The ordered p-values are denoted by $p_{(i)}$, $i = 1, \dots, s$, where $p_{(1)}$ is the smallest. The FWER is the probability of one or more false rejection in the 'family', so that FWER $\leq \alpha$ controls the FWER at level $\alpha = 0.05$. The Bonferroni correction (Bonferroni, 1936) achieves this control by rejecting H_i if $p_i \leq \alpha/s$, but may be quite conservative, when s is 'large', or if the p-values are highly positively correlated.[7] For example, if $s = 100$, p_i, has to be smaller than 0.0005 to reject the null at level α. When the set of hypotheses under consideration, H_i, $i = 1, \dots, s$, consists of hypotheses which are mostly true, few true null hypotheses would be rejected — the desired outcome. However, when the set is largely populated by false hypotheses, the Bonferroni correction results in low power (or high type 2 error) and the likely failure to reject false hypotheses.

A more powerful procedure is available if we are willing to allow more than one false rejection. Letting k-FWER denote the probability of rejecting k or more true null hypotheses, then controlling k-FWER at level α can be achieved by rejecting H_i if $p_i \leq k\alpha/s$ (see Lehmann and Romano, 2005). (1-FWER is of course FWER.)

More power can be achieved while maintaining control of the FWER or k -FWER by using 'stepdown' procedures in place of the onestep comparison of p_i to α/s for all i (or p_i to $k\alpha/s$ for all i for k-FWER). The stepdown procedure suggested by Holm (1979) generalizes Bonferroni by rejecting $H_{(i)}$, for $i = 1, \dots, s$, provided:

$$p_{(i)} \leq \alpha_i, \qquad \text{where } \alpha_i = \alpha/(s - i + 1), \tag{1}$$

and $H_{(1)}, \dots, H_{(i-1)}$ have all been rejected.[8] No hypotheses are rejected if $p_{(1)} > \alpha_1 = \alpha/s$.

Similarly, for k-FWER, Lehmann and Romano (2005) suggest a Holm-type stepdown procedure aimed at boosting the power. Choose the largest r such that the following inequalities hold:

$$p_{(1)} \leq \alpha_1, \dots, p_{(r)} \leq \alpha_r \tag{2}$$

where $\alpha_i = k\alpha/s$, for $i \leq k$, and $\alpha_i = k\alpha/(s + k - i)$, for $i > k$. Reject the set of hypotheses $H_{(1)}, \dots, H_{(r)}$. In the event that $p_{(1)} > \alpha_1$, no hypotheses are rejected.

The Holm procedure improves the ability to reject false hypotheses, and k -FWER control improves power at the cost of allowing k ($k > 1$) or more true nulls to be rejected. Nevertheless, it has been argued that the (Holm, 1979) stepdown procedure is only a little less conservative than Bonferroni (see, e.g., Efron and Hastie, 2016, p. 284–5, and Romano and Wolf, 2005). We consider whether these refinements make much difference to the application of MT adjustment to survey expectations.

Lehmann and Romano (2005) show that the onestep and stepdown procedures control k -FWER without requiring any restrictions on the admissible dependence structures for the p-values. While this might appear to be a desirable property of these approaches, Romano and Wolf (2005) note the conservativeness of these approaches stems from a failure to take into account the dependence structure.[9] Romano and Wolf (2005) propose a stepwise multiple testing strategy that accounts for the underlying dependence strategy, and improves on the approaches reviewed here. The Romano and Wolf (2005) strategy extends the 'reality

[6] The three case studies we consider are representative of a wider literature on the specific issues addressed. Many other studies address related issues, but sometimes consider aggregate quantities, such as consensus forecasts, where MT issues do not arise.

[7] The correction is based on the Bonferroni inequality $P\left(A_1, A_2, \dots, A_s\right) \geq 1 - \sum_{i=1}^{s} P\left(\overline{A}_i\right)$, where A_i is the event that test A_i does not reject, that is, $|t_i| < t_{\delta/2}$, where t_i is the test statistic and $t_{\delta/2}$ the critical value for a δ-level test. $\overline{A}_i$ is the complement of A_i, so that $P\left(\overline{A}_i\right) = \delta$. The inequality implies that $P\left(A_1, A_2, \dots, A_s\right) \geq 1 - s\delta$, so that setting $\delta = \alpha/s$ results in an FWER (at least one A_1 is false) of $1 - P\left(A_1, A_2, \dots, A_s\right) \leq \alpha$.

[8] Simes (1986) also provides a modification of the Bonferroni test procedure, but this is a test of $H_0 = \bigcap_{i=1}^{s} H_{0i}$ (the intersection or joint null hypothesis, that all hypotheses are jointly true), rather than of the individual H_{0i}'s. Cheng and Sheng (2017) consider testing joint null hypotheses and provide a review of some classical methods, but these are not our focus of interest. We suppose that some H_i's might be true and others false, and wish to discover how many are true and how many false.

[9] Romano and Wolf (2005) explain that "Loosely speaking, they achieve control of the FWE by assuming a worst-case dependence structure" (p. 1244) and to illustrate, note that if there were perfect dependence amongst the p-values, that is, identical p-values, rejecting H_i if $p_i \leq \alpha$ would control FWER, and rejecting H_i if $p_i \leq \alpha/s$ would be too conservative.

check' approach of White (2000), but as in White (2000), requires bootstrapping. The missing values in the series of forecasts for the individual respondents, and that the respondents may have been active participants at different times, would complicate the application of approaches that require bootstrapping, and we choose to avoid these.

When an allowance is made for MT, whether we reject a particular hypothesis, H_i, may depend on the number of other hypotheses we consider, and on the results we obtain for the other hypotheses. What is included in the 'family' becomes relevant when such adjustments are entertained (see Section 4.1 for an illustration). The role of the number of hypotheses in the 'family' (i.e., s) is obvious for the Bonferroni correction.

2.2. FDR control

The false discovery proportion (FDP) is the number of false rejections (F) divided by the total number of rejections (R). The FDP is defined as being zero when the denominator $R = 0$. Control of the FDP supposes one is willing to tolerate an increasing number of false rejections (i.e., 'false discoveries', F) as the number of rejections R increases. That is, false rejections are permitted provided these are made at an acceptably low rate relative to the number of discoveries. *Expected* FDP (that is, FDR) control at level q requires that $E(FDP) \leq q$. This can be achieved using a simple algorithm due to Benjamini and Hochberg (1995), referred to as BH henceforth.

For the ordered p-values, define $i_{\max}$ as the largest i for which:

$$p_{(i)} \leq \frac{i}{s} q, \tag{3}$$

where q is the desired FDR. Then reject $H_{(i)}$ for $i \leq i_{\max}$. (If $p_{(i)} > \frac{i}{s} q$ for all i, reject no null hypotheses). BH prove that this method is valid under the assumption that the p-values are independent. This method results in FDR being equal to $\pi_0 q$, where π_0 is the proportion of true null hypotheses, $N_0/s \leq 1$, where $N_0 = s - F$ is the number of true null hypotheses. Hence $FDR \leq q$. When $\pi_0 \ll 1$, the FDR is markedly lower than q. Methods of estimating π_0 are discussed in the next section.

A variant of FDR control suggested by Benjamini and Yekutieli (2001) (BY) allows for some forms of dependence in the p-values. BY replaces (3) by:

$$p_{(i)} \leq \frac{i}{s \times C_s} q \tag{4}$$

where $C_s = \sum_{i=1}^{s} \frac{1}{i}$. BY is obviously stricter than BH: the inequality is satisfied by hypotheses with lower p-values.

Rather than controlling the *expected* FDP, the FDR, Lehmann and Romano (2005) suggest controlling the FDP, in the sense that:

$$P\{FDP > \gamma\} \leq \alpha. \tag{5}$$

This can be achieved with a stepdown procedure, that compares the $p_{(i)}$ to α_i defined by:

$$\alpha_i = \frac{([\gamma i] + 1)\alpha}{s + [\gamma i] + 1 - i}. \tag{6}$$

(Here, $[z]$ is the greatest integer less than or equal to z.)

As above, if $p_{(1)} > \alpha_1$, no hypotheses are rejected. Otherwise, r is the largest value that satisfies (2) with α_i defined by (6), and the set of hypotheses $H_{(1)}, \ldots, H_{(r)}$ are rejected. Lehmann and Romano (2005) show that this procedure satisfies (5) with only relatively weak restrictions on the admissible dependence structures on the p-values.

Methods which control the FDP or FDR will be more liberal than controlling FWER. For example, from (3) it follows immediately that BH will likely generate more rejections ('discoveries') than the Bonferroni correction. If we suppose $q = \alpha$, then the Bonferroni-adjusted threshold for rejection is smaller than the FDR threshold for all but the smallest p-value hypothesis ($p_{(1)}$) since:

$$\frac{\alpha}{s} < \frac{i\alpha}{s} \text{ for } i = 2, \ldots, s \tag{7}$$

and is the same for $i = 1$.

Under FDR control, the dependence of the finding for H_i on the tests of the other hypotheses is readily apparent, because the position of i in the set of ordered p-values will be higher the smaller the other p_j's. Hence p_i will be compared to $\frac{k}{s} q$ with a higher k (where p_i is the kth smallest p-value). The greater the 'effect sizes' (equivalently, the smaller p-values) of the other hypotheses the more likely H_i will be rejected. In this sense FDR control depends on the evidence against the null of all the hypotheses in the family.

2.2.1. *q-values and estimating the proportion of true null hypotheses*

Given the $\{p_{(i)}\}$, we can estimate 'q-values' for each of the hypotheses. The q-value for hypothesis i is the expected proportion of false positives when hypothesis i is deemed significant. Note the q-value is not the probability that hypothesis i is a false positive. It is the minimum possible FDR at which hypothesis i is rejected. This mirrors the p-value - the minimum possible false positive rate at which we reject the null.

Storey and Tibshirani (2003) provide the following simple algorithm for calculating q -values. For the s ordered hypotheses, $p_{(1)} \leq p_{(2)} \cdots \leq p_{(s)}$, the q-value for the largest p-value hypothesis is:

$$\hat{q}\left(p_{(s)}\right) = \hat{\pi}_0 p_{(s)} \tag{8}$$

where $\hat{\pi}_0$ is an estimate of π_0, the proportion of true null hypotheses. Then for $i = s-1, s-2, \ldots, 1$:

$$\hat{q}\left(p_{(i)}\right) = \min\left(\frac{\hat{\pi}_0 s p_{(i)}}{i}, \hat{q}\left(p_{(i+1)}\right)\right) \tag{9}$$

where $\hat{q}\left(p_{(i)}\right)$ is the estimated q -value for the ith most significant (smallest p-value hypothesis).

The conservative approach is to assume $\pi_0 = 1$. When we set $\hat{\pi}_0 = 1$, then the algorithm set out in (8) to (9) is equivalent to the rule (3) of BH. That is, if we reject hypotheses for which the application of (8) to (9) result in $\hat{q}\left(p_{(i)}\right) < 0.05$, the set of rejections will exactly match BH control with $q = 0.05$.

Storey and Tibshirani (2003) propose an algorithm for estimating $\hat{\pi}_0$. For a given λ, Storey and Tibshirani (2003) suggest calculating:

$$\hat{\pi}_0(\lambda) = \frac{\sum_i^s 1_{(p_i > \lambda)}}{N} \times \frac{1}{1-\lambda} \tag{10}$$

for a grid of values $\lambda = 0, 0.01, 0.02 \ldots, 0.95$, and then fitting a natural cubic spline with 3 degrees of freedom of $\hat{\pi}_0(\lambda)$ on λ, denoted $\hat{f}$, and estimating π_0 by $\hat{f}$ $(\lambda = 1)$. Barras et al. (2010) (see also Glickman et al., 2014) find that simply setting $\lambda = \frac{1}{2}$ say works reasonably well, and little is gained by selecting over a grid of values for λ.[10]

q-values can be calculated with $\hat{\pi}_0 < 1$, as suggested by (8) to (9), or one could equivalently adjust the BH rule to:

$$p_{(i)} \leq \frac{i}{s\hat{\pi}_0} q \tag{11}$$

where $s\pi_0 = s_0$, so that relative to BH, (11) replaces s by an estimate of s_0, the number of true null hypotheses.

We use $\hat{\pi}_0 = 1$ in our empirical work, but indicate how the results would change if for example we set $\hat{\pi}_0 = \frac{1}{2}$. As noted above, using a unit value is the conservative assumption for FDR, and this allows a fair comparison with the Bonferroni correction.

2.2.2. Bayesian interpretation

It is illuminating that FDR can be given a Bayesian interpretation, and thus stands in stark contrast to the control of the type 1 error rate (or the extension to FWER) of classical frequentist hypotheses testing: see Efron and Hastie (2016) (section 15.3) for details.

Here we simply note that BH control amounts to rejecting H_i when the empirical Bayes posterior probability of hypothesis i being null is 'too small' (given that we have observed the 'effect size' and corresponding p -value), where by 'too small' is meant that it is less than $\pi_0 q$.

3. Monte Carlo

We carry out a Monte Carlo to provide evidence on the reliability of the testing procedures in Section 2 applied in the context of macro survey expectations. Specifically, we consider the MT strategies for the numbers of forecasters and forecast sample sizes typical of those available in surveys of macroeconomic expectations. The Monte Carlo aligns with the first of the empirical studies of survey expectations: the study of forecast efficiency and the reaction to news in Section 4.1. Section 4.1 provides additional discussion and motivation for the formulation below.

We consider the performance of the approaches when the null is true for all respondents, and when it is false for all respondents.

3.1. Data generation process and forecaster behaviour

The data generation process loosely matches U.S. quarterly output growth:

$$y_t = \beta_0 + \beta y_{t-1} + \eta_t \tag{12}$$

where η_t is an iid Gaussian innovation, $\eta_t \sim N\left(0, \sigma_\eta^2\right)$. Each agent is assumed to receive a noisy signal on the state of the economy y_t, give by:

$$s_{it} = y_t + \varepsilon_{it}, \tag{13}$$

where $\varepsilon_{it} \sim N\left(0, \sigma_{\varepsilon_i}^2\right)$, and $\sigma_{\varepsilon_i}^2 = \sigma_\varepsilon^2 \ \forall_i$ assuming homogeneity. Agent i's information set at time t, $\mathcal{I}_{i,t} = \left\{s_{it}, s_{it,t-1}, \ldots\right\}$ comprises the history of signals received by agent i through t, with past values of y unobserved. The optimal forecast of t, $f_{it|t}$, incorporates s_{it} via:

$$\begin{aligned} f_{it|t} &= K s_{it} + (1-K) f_{it|t-1} \\ &= f_{it|t-1} + K\left(s_{it} - f_{it|t-1}\right), \end{aligned} \tag{14}$$

[10] The motivation for this approach is that true null p-values are $U(0,1)$. The majority of p-values larger than a high enough threshold λ, say $\lambda = \frac{1}{2}$, are for true H_i. Eq. (10) calculates the proportion of such forecasters, and then scales this (by the $(1-\lambda)^{-1}$ factor) over the whole region between 0 and 1.

This approach has been proposed in the context of genomewide studies where thousands of genes (i.e., null hypotheses) are being considered.

where K is the Kalman gain. Eq. (14) updates the forecast of t based on information through $t-1$, $f_{it|t-1}$, using the optimal (in a Minimum Mean Squared Error sense) weight K, where $K = \Sigma/\left(\Sigma + \sigma_\varepsilon^2\right)$, and:

$$\Sigma = \frac{1}{2}\left(-\left(1-\beta^2\right)\sigma_\varepsilon^2 + \sigma_\eta^2 + \sqrt{\left[\left(1-\beta^2\right)\sigma_\varepsilon^2 - \sigma_\eta^2\right]^2 + 4\sigma_\varepsilon^2\sigma_\eta^2}\right) \tag{15}$$

(see, e.g., Bordalo et al., 2020). The 1-step forecast is given by $f_{i,t+1|t} = \beta_0 + \beta f_{it|t}$, and so on.

We report tests for the two hypotheses considered in Section 4.1: the test of forecast efficiency, and the optimal reaction to news. The first is based on estimating for each individual:

$$y_{t+h} - f_{i,t+h|t} = \alpha_i + \beta_i f_{i,t+1|t} + u_{i,t+h}. \tag{16}$$

The null is the simple hypothesis H_i: $\beta_i = 0$. (We consider only the slope, and leave the intercept unrestricted.) We consider only $h = 1$.

The test of the reaction to news is based on:

$$y_{t+h} - f_{i,t+h|t} = \alpha_i + \beta_i\left(f_{i,t+h|t} - f_{i,t+h|t-1}\right) + u_{i,t+h} \tag{17}$$

and the null is again H_i: $\beta_i = 0$, with $h = 1$.

For forecasts generated by Eqs. (13) to (15), H_{0i} is true both when we test for forecast efficiency and for the optimal reaction to news. (See Section 4.1 for further discussion).

We model the departure from optimality by assuming Diagnostic Expectations (DE), as in Bordalo et al. (2020), although another possibility would be to allow incorrect values of the law of motion parameters, β_0 and β, in (12). Under DE, (14) becomes:

$$f_{it|t} = f_{it|t-1} + (1+\theta)K\left(s_{it} - f_{it|t-1}\right). \tag{18}$$

where $\theta > 0$ indicates news is over-weighted relative to the optimal amount given by the Kalman gain K.

Forecasters are assumed to know the parameters β_0, β, and the variances of the disturbances and signals, and hence K.

The model is loosely calibrated on U.S. real quarterly GDP growth. We suppose $\beta_0 = 0.50$, and $\beta = 0.36$. This reproduces the unconditional growth rate of quarterly real GDP of 0.78 for the period 1947:1 – 2018:2 (2018:3 data vintage). The AR(1) model estimated standard error is $\sigma_\eta = 0.88$. These values are used for β_0, β and σ_η throughout.

We set $\sigma_\varepsilon = \frac{1}{2}$ and 3, and assume homogeneous forecasters for simplicity, so that $\sigma_{\varepsilon_i}^2 = \sigma_\varepsilon^2$ for all i. For the higher value of σ_ε the forecasters' signals are less informative, and their forecasts less accurate, other things being equal.

Under the alternative (that is, under DE), we set $\theta = \frac{1}{2}$ in (18).

3.2. Simulation results

The number of replications is 10,000. We consider $T = 25, 50$, which affects the estimation of the t-statistic/p-value for H_i. We set $N = 50$ throughout. $N = 50$ is typical of the number of respondents to the U.S. SPF who have responded sufficiently often to provide a useable number of forecasts.

Table 1 reports the results for the test of forecast efficiency. We consider BH and BY, and Bonferroni, but omit the other FWER control techniques to save space. Table 2 has the results for the test of the reaction to news.

Consider first Table 1, Panel A, $\pi_0 = 1$, when all the null hypotheses are true. The results show that the liberal strategy (making no allowance for MT) fails to control false discoveries (column 3), and the FWER (column 4). The proportion of replications for which the ratio of false discoveries to total rejections is less than 5% is lower than 25% when $\sigma_\varepsilon = 3$, for both values of T. For this value of σ_ε, the FWER (the proportion of replications for which the true null is rejected for one or more forecasters) is around 98%. BY is generally superior to BH in terms of controlling the FDP on around 99% of the replications, and resulting in an FWER of less than 10% when T takes on the larger value ($T = 50$). Bonferroni works well when $\sigma_\varepsilon = \frac{1}{2}$, but less well when $\sigma_\varepsilon = 3$.

Panel B $\pi_0 = 0$ shows the results for the other extreme when H_i is false for all forecasters. Now FDP control is achieved for all approaches by construction, because all rejections are true discoveries. (Similarly for FWER control. All rejections are correct.) The rejection rate by Bonferroni is much lower when $\sigma_\varepsilon = \frac{1}{2}$. The lower value of σ_ε corresponds to positively-correlated p-values across tests, because the signals are more informative and the forecasts are closer to the true values. The positive correlation accounts for the lower power of Bonferroni. However, BY is also low for this value of σ_ε. When $\sigma_\varepsilon = 3$ BY clearly outperforms Bonferroni.

The results in Table 2 for the test based on (17) are qualitatively similar.

To briefly summarize the findings of the Monte Carlo: BY works reasonably well for the relatively small samples of forecasters ($N = 50$) and numbers of forecasts ($T = 25, 50$) typically available in macro-surveys. BY outperforms the Liberal approach – when $\pi_0 = 1$ – by controlling 'false discoveries', and rejects one or more true nulls (FWER) for a similar proportion of replications as Bonferroni. When $\pi_0 = 0$, BY exhibits greater power than Bonferroni. However, Bonferroni is not markedly less powerful than BY (the preferred FDR technique) in the setups we consider, and this is also evident in some of our empirical results in Section 4 where BY and Bonferroni deliver similar numbers of rejections.

Table 1
Noisy information DGP, Test of forecast efficiency.

	Liberal			BH			BY			Bonferroni		
	2 Rej	3 FD	4 FWER	5 Rej	6 FD	7 FWER	8 Rej	9 FD	10 FWER	11 Rej	12 FD	13 FWER
Panel A. $\pi_0 = 1$												
25, $\frac{1}{2}$	0.071	0.750	0.409	0.020	0.957	0.061	0.006	0.985	0.025	0.004	0.982	0.052
50, $\frac{1}{2}$	0.065	0.776	0.370	0.017	0.968	0.042	0.004	0.989	0.016	0.003	0.987	0.034
25, 3	0.094	0.180	0.987	0.010	0.953	0.309	0.003	0.995	0.141	0.007	0.990	0.295
50, 3	0.088	0.231	0.981	0.007	0.973	0.232	0.002	0.997	0.094	0.005	0.995	0.222
Panel B. $\pi_0 = 0$												
25, $\frac{1}{2}$	0.254	1	0	0.140	1	0	0.057	1	0	0.034	1	0
50, $\frac{1}{2}$	0.369	1	0	0.233	1	0	0.103	1	0	0.059	1	0
25, 3	0.842	1	0	0.807	1	0	0.638	1	0	0.468	1	0
50, 3	0.980	1	0	0.978	1	0	0.932	1	0	0.818	1	0

For each method we show the average (across replications) proportion of rejections, 'Rej'; the proportion of replications for which the proportion of false discoveries was less than 0.05, 'FD'; and the proportion of replications for which one or more true nulls were rejected, 'FWER'.
a, b in the first column refer to the sample size T and σ_ϵ.

Table 2
Noisy information DGP, Test of reaction to new information.

	Liberal			BH			BY			Bonferroni		
	2 Rej	3 FD	4 FWER	5 Rej	6 FD	7 FWER	8 Rej	9 FD	10 FWER	11 Rej	12 FD	13 FWER
Panel A. $\pi_0 = 1$												
25, $\frac{1}{2}$	0.056	0.785	0.397	0.014	0.970	0.044	0.004	0.988	0.018	0.003	0.986	0.037
50, $\frac{1}{2}$	0.052	0.799	0.376	0.012	0.976	0.032	0.002	0.993	0.011	0.002	0.992	0.025
25, 3	0.063	0.428	0.943	0.004	0.990	0.149	0.001	0.999	0.059	0.003	0.999	0.143
50, 3	0.056	0.508	0.920	0.002	0.996	0.090	0.001	1.000	0.028	0.002	1.000	0.088
Panel B. $\pi_0 = 0$												
25, $\frac{1}{2}$	0.213	1	0	0.105	1	0	0.040	1	0	0.024	1	0
50, $\frac{1}{2}$	0.344	1	0	0.204	1	0	0.081	1	0	0.046	1	0
25, 3	0.807	1	0	0.754	1	0	0.529	1	0	0.357	1	0
50, 3	0.978	1	0	0.976	1	0	0.914	1	0	0.759	1	0

For each method we show the average (across replications) proportion of rejections, 'Rej'; the proportion of replications for which the proportion of false discoveries was less than 0.05, 'FD'; and the proportion of replications for which one or more true nulls were rejected, 'FWER'.
a, b in the first column refer to the sample size T and σ_ϵ.

4. Survey expectations applications

The illustrations in this section use the U.S. Survey of Professional Forecasters (SPF).[11] The SPF is a quarterly survey of macroeconomic forecasters of the U.S. economy that began in 1968, originally run by the American Statistical Association (ASA) and the National Bureau of Economic Research (NBER), and since June 1990 by the Philadelphia Fed (see Croushore, 1993). A recent survey of professional forecasters' expectations by Clements et al. (2023) discusses the SPF in some detail, reviews the types of analyses which have been conducted, and some of the findings, concluding that individual-level forecaster heterogeneity is a key feature of the forecasts.

4.1. Weak efficiency tests and forecast error — forecast revision regressions

Much of the recent literature on survey expectations seeks to explain forecaster disagreement in terms of theories of informational rigidities and 'rational inattention'[12], and the responsiveness of expectations to news, and this motivates our first two sets of

[11] https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/survey-of-professional-forecasters

[12] On sticky information, see *inter alia* Mankiw and Reis (2002) and Mankiw et al. (2003), and Coibion and Gorodnichenko (2012), and on noisy information, Woodford (2002), Sims (2003) and Coibion and Gorodnichenko (2012), *inter alia*.

behavioural hypotheses. The Monte Carlo in Section 3 tests for the responsiveness to news (and the closely-related notion of forecaster efficiency) assuming noisy information.

In essence, under dispersed noisy information, agents receive noisy signals, as described in Section 3, but provided they update optimally, their forecast errors will be unrelated with their forecast revisions and their forecasts. Intuitively, each forecaster makes optimal use of their signal, and the forecast error is not systematically related to the forecast revision (or forecast). Each forecaster correctly downweights his/her information because it is noisy, but because the private noise cancels in the aggregate, the average forecast under-responds to the new information.[13]

Our interest is in testing $\alpha_i = 0$ and $\beta_i = 0$ for $i = 1, \dots, s$, in Eqs. (16) and (17). As we wish to allow for the possibility that H_i is true for a set of forecasters, but false for the remainder, pooled or fixed effects panel regressions (where the slope is the same across all respondents) are not appropriate. Both Broer and Kohlhas (2021) and Bordalo et al. (2020) apply the test based on (17), and find forecasters are over-confident, in the sense that they *over*-react to new information: β is found to be negative. The test based on (16) is the test of weak-efficiency of Mincer and Zarnowitz (1969), as reported by Clements (2022), Table 2. When the null is true, the forecast is efficient in the sense that the resulting forecast error is not systematically related to the forecast. Clements (2022) shows that forecast efficiency can hold without the forecaster i making use of all relevant information, and in the presence of private information.[14]

For both tests the right-hand-side actual values $\{y_{t+h}\}$ are the advance estimates, and inference is based on a HAC estimator of the variance–covariance matrix.

We consider the impact of allowing for MT for both tests.

To match the sample period and variables used by Clements (2022), we use the SPF multi-horizon forecasts of real GDP, consumption, and investment from 1990:4 to 2017:2 (that is, from when the SPF was administered by the Philadelphia Fed). We consider the $s = 50$ individuals who made the most forecasts during this period. The average number of forecasts per person for this group was 55 (for each variable and at each forecast horizon), with a minimum and maximum of 31 and 98 attesting to the large number of missing forecasts for some respondents.

Since we use the point forecasts, a maintained assumption throughout is that the point forecasts are the means of the respondents' subjective distributions.[15]

4.1.1. Empirical findings

Consider firstly the weak-efficiency (Mincer and Zarnowitz, 1969) test results reported by Clements (2022), Table 2. The null hypothesis is that $\alpha_i = 0$ and $\beta_i = 0$, and the test statistic uses a HAC estimator of the variance–covariance matrix. When no adjustment is made for multiple testing, the row for 'Liberal' in Table 3, Panel A shows that using a 5% significance level, the null is rejected for 40% or more of the U.S. SPF respondents' current-quarter ($h = 0$) forecasts of consumption, investment and GDP growth. For their four quarter ahead ($h = 4$) forecasts, the null was rejected for more than three-quarters of respondents (and for all but 3 of the 50 respondents for GDP growth). The Bonferroni correction reduces the significance level to $0.05/50$ for each combination of variable and forecast horizon (when we have 50 respondents). The Bonferroni correction results in the null only being rejected for around 1 in 6 respondents for the short-horizon forecasts. The proportions of rejections at the longer horizon are reduced but remain above a half of all respondents. The Holm procedure is expected to be more powerful than Bonferroni, while still controlling the FWER (here, at the 5% level), but in our case its application yields very similar results to Bonferroni.

BH control at $q = 0.05$ has no effect on the proportion of respondents for whom we reject at the longer horizon, but reduces the rejections at the short horizon. However, as noted in Section 2, BH FDR control requires independence of the p-values, which may not hold. BY does not require independence, and consequently is stricter than BH. The use of BY approximately halves the number of rejections for the short-horizon forecasts, compared to BH, but at the longer horizon has a more benign effect, where the proportion of rejections remains above two thirds for all 3 variables. In fact short-horizon rejections are markedly reduced for all methods of correcting for MT other than BH. Holm k -FWER with $k = 2$ generates only a small additional number of rejections relative to when $k = 1$ (Holm, in the table), and is similar to BY.

Finally, if we were to use an estimated value of π_0 in an MT adjustment strategy ((e.g., (11))), we would be able to control the FDR at 5% while rejecting a greater number of null hypotheses, given that the estimated p-values suggest a π_0 well below one. If $\pi_0 = 1$ the distribution of the p-values would be roughly uniform on the unit interval. The histogram of the p-values for the current-quarter consumption growth forecast MZ tests is shown in Fig. 1, and that for the four-quarter ahead consumption growth forecasts in Fig. 2.[16] Both figures confirm the excess number of low p-values, relative to a uniform, casting doubt on $\pi_0 = 1$ for both horizons. This suggests that setting $\pi_0 = 1$ is conservative, especially for the short-horizon forecasts, but we do not explicitly consider other values of π_0.[17]

[13] Hence for *aggregate* forecast quantities, the regression of the aggregate forecast error on the aggregate forecast revision, between origins $t-1$ and t, say, will result in a positive coefficient, related to K, given by $(1-K)/K > 0$ (see, e.g., Coibion and Gorodnichenko, 2015). That is, aggregate forecasts under-react, resulting in a positive correlation between the error and the revision.

[14] Fuhrer (2018) argues that the forecast rather than the revision to the forecast has the greater predictive power for the forecast error in most cases, and suggests this weakens the interpretation that there is over-response to the 'news' embodied in the revision.

[15] This need not be the case. The SPF does not specify that respondents should report their means, and they may report other measures of central tendency, and what they report may vary over forecast rounds. However, the interpretation of the point forecasts as mean values is a commonly-made assumption. There are a number of papers that compare the respondents histogram forecasts and point forecasts (e.g., Engelberg et al., 2009 and Clements, 2009).

[16] These are broadly representative of the findings for investment and output growth, which are not shown to save space.

[17] In the Appendix we provide details of the application of the strategies for each individual in Tables 6 and 7, and illustrate how the findings change for other values of π_0.

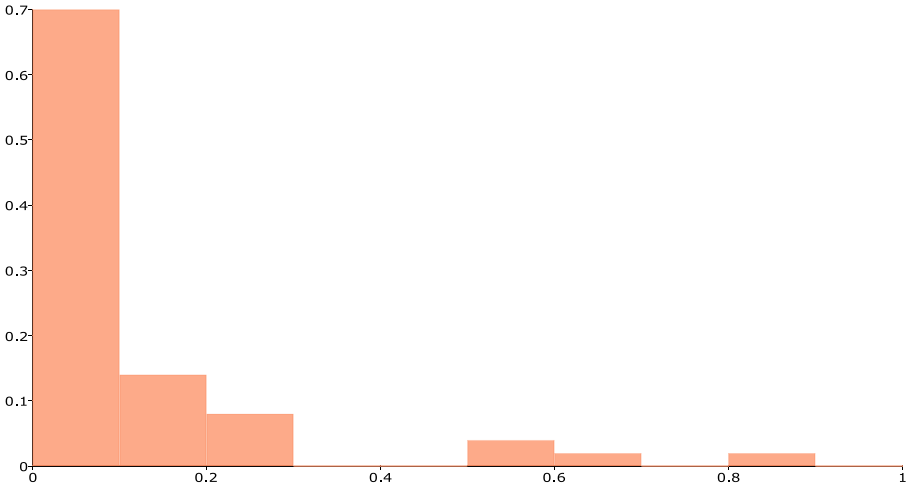

Fig. 1. *p*-values for tests of the MZ efficiency hypothesis for each respondent's current-quarter consumption growth forecasts.

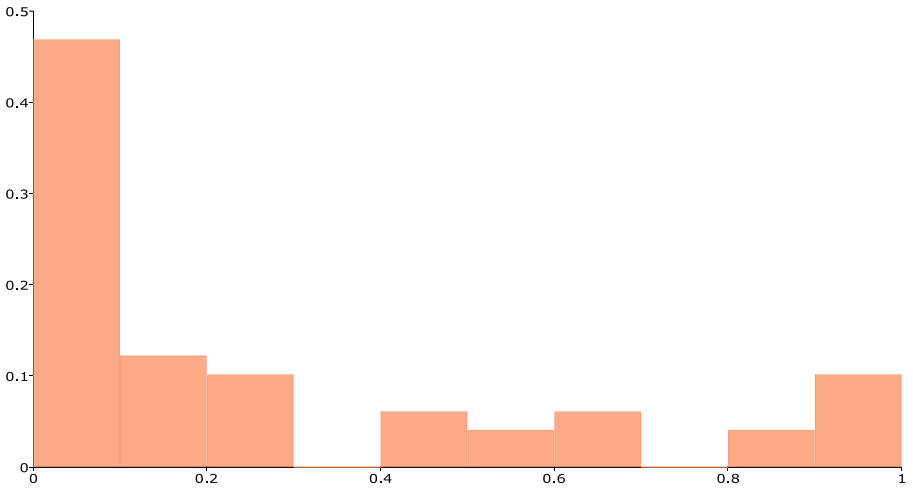

Fig. 2. *p*-values for tests of the MZ efficiency hypothesis for each respondent's 4-quarter ahead consumption growth forecasts.

We also consider the impact of MT on the optimal response to news hypothesis, based on the forecast error — forecast revision regression of ((17)). Table 3, Panel B shows that the Liberal strategy leads to fewer rejections than for weak-efficiency (MZ) tests, especially at $h = 4$, where there are only half as many. This is consistent with the findings of Fuhrer (2018). However, to what extent is this due to neglecting MT issues? The Bonferroni correction approximately halves the $h = 4$ respondent rejection rates: a much larger impact than for the MZ tests. In addition, the Bonferroni rejection rates are little more than 1 in 10 at most. As for the MZ efficiency tests, more powerful tests than Bonferroni generally result in similarly low numbers of rejection. Only FDR control by BH has a muted effect relative to the Liberal approach.

In summary, we find that controlling for MT results in a sizeable reduction in the number of rejections for the over-reaction hypothesis, and for the short-horizon MZ hypotheses, but not for the long-horizon MZ hypotheses. With the exception of FDR control by BH, it makes little difference whether we control the FDR or the FWER, and which variant of these two types of control is used. We have assumed that hypotheses regarding short and long-horizon forecast behaviour belong in different families when we consider multiple testing. This seems reasonable *ex post* given that the test results for the two seem somewhat different. Moreover, there are theoretical reasons to suppose forecaster-behaviour at the two horizons may differ. Short-term forecasting or nowcasting may be accomplished by projecting current trends and considering recent indicators, whereas forecasting a year-ahead may require more expertise, and possibly an economic model. Hence we consider hypotheses related to short and long-horizon forecast performance

Table 3
Multiple testing and individual forecast efficiency and over-reaction regressions.

	Consumption		Investment		Output	
	$h = 0$	$h = 4$	$h = 0$	$h = 4$	$h = 0$	$h = 4$
Panel A. Forecast Efficiency Regression						
Liberal	0.56	0.76	0.40	0.90	0.42	0.94
BH	0.44	0.76	0.32	0.90	0.32	0.94
BY	0.20	0.68	0.18	0.76	0.20	0.82
FDP	0.16	0.66	0.18	0.76	0.14	0.92
Bonferroni	0.16	0.52	0.16	0.62	0.14	0.72
Holm	0.16	0.56	0.18	0.66	0.14	0.74
Holm k-FWER	0.20	0.66	0.20	0.76	0.20	0.82
Panel B. Over-reaction Regression						
Liberal	0.52	0.37	0.32	0.55	0.26	0.54
BH	0.30	0.24	0.10	0.45	0.22	0.48
BY	0.16	0.16	0.04	0.33	0.10	0.25
FDP	0.16	0.12	0.04	0.29	0.10	0.23
Bonferroni	0.16	0.12	0.04	0.29	0.10	0.23
Holm	0.16	0.12	0.04	0.29	0.10	0.23
Holm k-FWER	0.18	0.18	0.08	0.33	0.12	0.25

The table shows the proportion of U.S. SPF respondents for whom the null is rejected at the 5% level when no allowance is made for multiple testing (Liberal), and for controlling the FWER or the FDR. BH is Benjamini and Hochberg (1995) FDR control at the 5% level. BY is Benjamini and Yekutieli (2001) control at the 5% level. FDP sets the probability that the false discovery proportion exceeds 5% at less than 5%. Bonferroni controls the FWER at 5%. Holm is a stepdown generalization of Bonferroni. Holm k-FWER is a stepdown implementation of k-FWER control (at the 5% level), with $k = 2$.
The figures for the liberal strategy for the forecast efficiency regression correspond to those in Clements (2022), Table 2, page 547.
The actual values are the Bureau of Economic Analysis advance estimates, as made available in the Real Time Data Set for Macroeconomists (RTDSM) maintained by the Federal Reserve Bank of Philadelphia: (Croushore and Stark, 2001).

Table 4
Multiple testing and individual forecasters' perceptions of the persistence of output growth.

	Annual	10 year
Liberal	0.8519	0.4815
BH	0.8519	0.2963
BY	0.7778	0.0741
FDP	0.8148	0.0741
Bonferroni	0.7778	0.0741
Holm	0.7778	0.0741
Holm k-FWER	0.8148	0.2222

The table shows the proportion of U.S. SPF respondents for whom the null is rejected at the 5% level when no allowance is made for multiple testing (Liberal), and for controlling the FWER or the FDR.
The figures for the liberal strategy correspond to those in Clements (2020), Table 1.
See notes to Table 3 for details of testing approaches, or the main text.

as constituting two different families of hypotheses, and do not explore the consequences of considering them as a single family here.[18]

4.2. Forecasters' perceptions of persistence

The second study we consider is the analysis of forecasters' perceptions of output growth persistence of Clements (2020). Following on from Krane (2011) and Bluedorn and Leigh (2018) and others, Clements (2020) investigates the beliefs or perceptions of professional forecasters regarding the persistence of shocks to output, and in particular whether they are believed to be permanent, or only have a temporary effect. We regress the revision to the long-horizon (10 year) average annual growth on the revision in the quarterly forecast growth rate at t:

$$r_t\left[\Delta y_{t,10}\right] = \alpha_i + \beta_{10,i} . r_t\left[\Delta y_{t,1}\right] + \upsilon_{i,t} \tag{19}$$

[18] A case could also be made for treating hypotheses regarding the forecasts of the different variables as members of the same family.

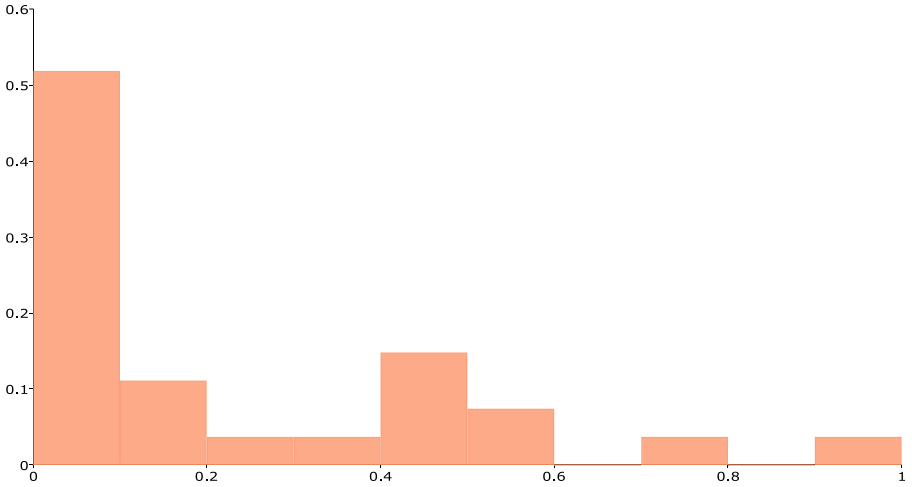

Fig. 3. *p*-values for tests of forecasters' perceptions of output persistence for each respondent, using 10-year growth forecasts.

where $r_t\left[\Delta y_{t,10}\right]$ is the revision in the 10-year average forecasts made in the first quarters of the years t and $t-1$, and $r_t\left[\Delta y_{t,t}\right]$ is the revision in the current-quarter growth rate (for the first quarter of the year) between the first quarters of the years t and $t-1$. The 10-year annual-average real GDP growth forecasts (SPF variable identifier RGDP10) were only collected for first quarter of the year surveys, accounting for the pattern of revisions we adopt.

We also estimate a regression which replaces $r_t\left[\Delta y_{t,10}\right]$ by the revision in the current-year annual growth rate $r_t\left[\Delta y_{t,a}\right]$ (between the first quarters of the years t and $t-1$). The right-hand-side variable is again the revision to the current-quarter growth rate between the same two forecast origins.

4.2.1. Empirical findings

Regressions are run for each individual, and are shown in Clements (2020, Table 1). The estimates of β are found to vary widely, from −0.09 to 0.19 for the 10-year annual average growth rate, and from 0.053 to 0.782 for the annual average. Just under a half of the twenty seven β estimates are significantly different from zero for the 10-year forecasts, at the 5% level, while the null that the slope coefficient is zero is rejected for 24 out of the 27 forecasters for the annual forecasts, again at the 5% level.

Clements (2020) considers whether the times of participation as a survey respondent help explain the cross-sectional differences in the estimates of the perceptions of persistence, and whether the differences can partly be accounted for by small-sample variability in the estimates. Our interest is instead whether the estimates of medium term and long-run (10-year) persistence hold up once an allowance is made for MT.

Table 4 reports the findings, and Fig. 3 and Fig. 4 plot the *p*-values for each individual for the 10-year and annual forecasts. Firstly, consider the annual forecasts in the first column. Making an allowance for multiple testing has little discernible effect. The conservative Bonferroni correction reduces the number of rejections from 23 (out of 27) to 21, while BH results in the same number of rejections of the null as for the Liberal approach. Fig. 4 plots the *p*-values. The large number of low values explains why controlling either FWER of FDR has little effect on the number of 'positive' results.

For the 10-year forecasts (see the second column of the table), the pattern is rather different. As is evident from Fig. 3, there are fewer low *p*-values compared to Fig. 4. The Bonferroni correction reduces the rejections from just under 1 in 2 to less than 1 in 10, and the other FWER control techniques have the same effect. The effect of FDR control is now more sensitive to whether we use BH or BY control. Siding with BY, to allow for dependence in the *p*-values, matches FWER control, leading to the overall conclusion of little evidence of long-run (10-year) persistence.

4.3. Forecasters' perceptions of uncertainty

Lastly, we re-analyse the study by Clements (2014) of the accuracy of individual forecasters' beliefs about the uncertainty they face. Clements (2014) considers whether professional forecasters tend to be under- or over-confident, and whether this depends on the forecast horizon. In summary, Clements (2014) finds that forecasters tend to over-confidence at horizons in excess of one year, but under-confidence at within-year horizons. A measure of 'perceived' or *ex ante* uncertainty – derived from the reported forecast

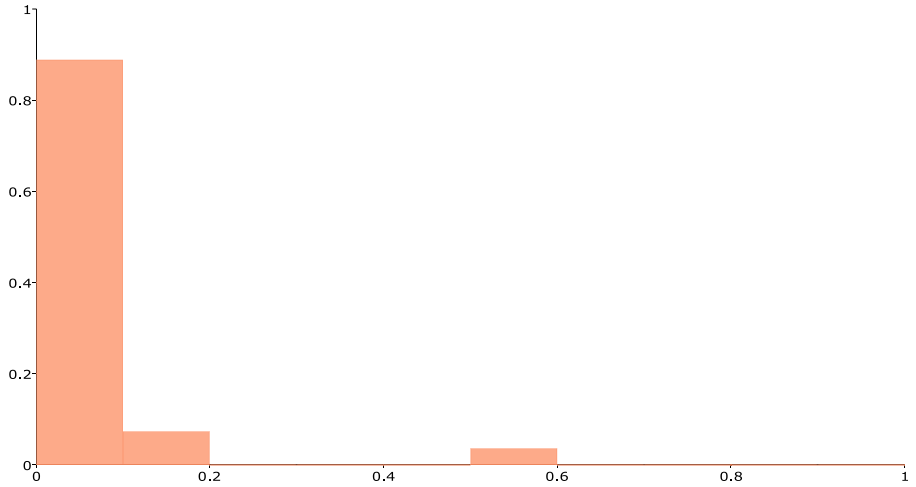

Fig. 4. *p*-values for tests of forecasters' perceptions of output persistence for each respondent, using annual growth forecasts.

histograms – remains at a high level compared to the *ex post* measure as the horizon shortens. These findings are again based on the U.S. SPF.[19] We consider whether the overall findings are affected by making an allowance for multiple testing.

The histograms underpinning the estimates of *ex ante* uncertainty refer to the annual change from the previous year to the year of the survey, as well as of the survey year to the following year. As an example, a Q1 survey will provide a 4-step ahead forecast of the current year's growth rate, and an 8-step ahead forecast of next years' growth rate, and a Q4 survey will provide a 1-step ahead forecast of the current year's growth rate, and a 5-step ahead forecast of next period's growth rates. This generates sequences of fixed-event histogram forecasts, with horizons from 8 down to 1 quarter ahead, for the annual growth rates of GDP and the GDP deflator inflation rate for each year. The histogram variance is the estimate of *ex ante* uncertainty. The variance can be calculated by assuming the probability mass is uniform within each of the histogram bins, or that it is located at the mid-points of the bins, or by fitting a parametric distribution. Clements (2014) follows Engelberg et al. (2009) and fits generalized beta distributions. Here we fit normal distributions, with similar results, and triangular distributions when probabilities are assigned to one or two histogram bins (following Engelberg et al., 2009). Clements (2014) uses the surveys from 1981:3 up to 2010:4. Point forecasts are provided for the same horizons and quantities (i.e., annual average growth rates) as the histograms, and these are used to construct forecast errors using real-time actual values.[20] The (squares of the) forecast errors are used to proxy *ex post* uncertainty, and are compared to *ex ante* uncertainty as described below.

4.3.1. Empirical findings

A formal test of whether a respondent's subjective assessments of uncertainty deviate systematically from their *ex post* uncertainty is constructed as follows. For respondent i, and horizon h, the *ex ante* $\sigma_{i,t|t-h}$ and *ex post* (based on the forecast error, $e_{i,t|t-h}$) uncertainty assessments are compared by calculating $w_{i,t|t-h} = e_{i,t|t-h}/\sigma_{i,t|t-h}$, and then testing the null $E\left(w^2_{i,t|t-h}\right) = 1$ using a two-sided alternative. Here t indexes years, and h the horizon, so e.g., $h = 1$ indicates a fourth quarter survey forecast of the current year, and $h = 5$ a fourth quarter survey forecast of the following year.

Following Clements (2014), Table 5 reports the results for all the within-year forecasts taken together ('1-4' in the table), and for all the next-year forecasts taken together ('5-8'). Within-year forecasts are of the current-year annual growth rates (relative to the previous year), and next-year forecasts are of the year after the survey quarter year, relative to the survey quarter year.

Tests are also run which adjust for potential bias in the point forecasts, by replacing $w_{i,t|t-h} = e_{i,t|t-h}/\sigma_{i,t|t-h}$ with $w_{i,t|t-h} = (e_{i,t|t-h} - e_{i,h})/\sigma_{i,t|t-h}$, where $e_{i,h}$ is the sample mean of the forecast errors.

When no adjustment is made for multiple testing, Table 5 shows that the null is rejected for around a quarter of all respondents for GDP growth (current and within-year), and for over a half of respondents for within-year inflation forecasts, and closer to a

[19] Using the same approach, Knüppel and Schultefrankenfeld (2019) report similar findings for the inflation uncertainty forecasts from the Bank of England, the Banco Central do Brasil, the Magyar Nemzeti Bank and the Sveriges Riksbank. That is, these central banks' uncertainty forecasts also tend to be underconfident at short horizons and overconfident at longer horizons.

[20] These are again taken from the Real Time Data Set for Macroeconomists(RTDSM) maintained by the Federal Reserve Bank of Philadelphia: (Croushore and Stark, 2001).

Table 5
Summary of tests of individuals — proportion of regressions for which we reject $E\left(w_{i,t|t-h}\right)^2 = 1$.

	NBA		BA		NBA		BA	
	1–4	5–8	1–4	5–8	1–4	5–8	1–4	5-8
	Output growth				Inflation			
Liberal	0.28	0.28	0.27	0.31	0.57	0.29	0.64	0.28
BH	0.21	0.00	0.24	0.03	0.55	0.08	0.63	0.13
BY	0.14	0.00	0.20	0.00	0.47	0.03	0.55	0.07
FDP	0.13	0.00	0.17	0.00	0.45	0.03	0.48	0.07
Bonferroni	0.13	0.00	0.16	0.00	0.41	0.03	0.43	0.07
Holm	0.13	0.00	0.17	0.00	0.41	0.03	0.45	0.07
Holm k-FWER	0.14	0.00	0.19	0.02	0.43	0.03	0.48	0.07

For a given forecast horizon, for each individual with a sufficient number of forecast observations, we regress either $w^2_{i,t|t-h}$ ('NBA' - No Bias Adjustment) or $[(e_{i,t|t-h} - e_{i,h})/\sigma_{i,t|t-h}]^2$ ('BA' - Bias Adjusted'), on a constant, and test the hypothesis that the constant is one. We report rejection rates for the 5% significance levels. We consider together all the within-year forecasts (denoted '1-4') and all the next-year forecasts (denoted '5-8').
The results in the first 4 columns replicate part of Clements (2014), Table 5, which ignores multiple testing issues. (The results are similar but not exactly the same. Here we fit Gaussian distributions to the histograms. Clements (2014) fits Generalized Beta distributions).
See notes to Table 3 for details of testing approaches, or the main text.

quarter for next-year inflation forecasts. (The within-year forecast rejections are mainly due to under-confidence, and the next-year due to over-confidence, but we only report two-sided tests here). These findings are not much affected by whether or not a bias adjustment is made.

What happens if we make an allowance for MT? For next-year forecasts of output growth, any of the corrections result in no rejections. The evidence against the null for the current-year forecasts is also reduced, with rejections for 1 in 6 or fewer, with little variation across the form of FDR or FWER control.

For inflation we observe a similar outcome for the longer-horizon forecasts: rejections are greatly reduced (to less than 1 in 10, apart from for the more liberal BH FDR control when a bias correction to the forecast errors has been applied). But MT leaves the null hypothesis rejections largely intact for the within-year inflation forecasts. BH results in a few more rejections than the other MT strategies, while BY and FDP, and the FWER strategies, all generate similar numbers of rejections.

We conclude that the null is rejected for nearly a half of the respondents at within-year horizons for inflation when an allowance is made for MT. For output growth, MT reduces the rejections of the null to 1 in 6 at the within-year horizons, and removes all evidence against the null at the longer horizons.

5. Conclusions

In health and medical research studies where a large number of null hypotheses are tested (e.g., in genomewide association studies), the potential inferential problems of multiple testing are apparent. Following on from Benjamini and Hochberg (1995), controlling the false discovery rate (FDR) has become increasingly popular, as an alternative to controlling the family-wise error rate (FWER).

We investigate the consequences of making an allowance for multiple testing for testing hypotheses about individual respondents' expectations. Typically far fewer hypotheses are run than in medical research or finance settings. Depending on the behavioural hypothesis of interest, for quarterly macro surveys of professional forecasters, such as the U.S. SPF, there may be 50 or fewer individuals to be tested. The results of a Monte Carlo study suggest that the relatively small numbers of hypotheses being tested (relative to the number in genomewide association studies) does not invalidate the use of FDR. We investigate whether controlling for the MT makes a material difference to the inferences we make about macro-expectations formation for a number of recent studies. We find that whether or not controlling for MT matters depends on the hypothesis being tested. However, one of our key findings is that the FWER and FDR approaches, with the exception of FDR-BH, tend to generate similar numbers of rejections. The BH approach is more liberal, and closer to the unadjusted rejected rates. BH requires the independence of the p -values, so may not be appropriate.

To the best of our knowledge there are no other papers that address multiple testing in the context of individuals' expectations using macro surveys.

Multiple testing adjustments are considered for three papers that investigate forecaster behaviour by testing hypotheses for individual forecasters. The first is Clements (2022), who considers whether forecasters are weakly efficient, in the sense of Mincer and Zarnowitz (1969). In addition to considering whether multiple testing considerations affect the findings for weak efficiency, the results for testing the closely related notion of the optimal response to news are also investigated. Without any adjustments, weak efficiency is rejected for nearly half the respondents for short horizon forecasts, and for more than three-quarters of respondents' year ahead forecasts. Bonferroni correction reduces the number of rejections to low proportions for the short-term forecasts. Other approaches to controlling the FWER that are expected to be less conservative deliver similar numbers of rejections to Bonferroni.

FDR-BH control at 5% has little effect on the number of rejections of efficiency, but allowing for dependence in the test outcomes using FDR-BY again reduces the proportions of rejections to low levels for the short-horizon forecasts. The longer-horizon forecasts are much less affected by MT adjustments, and the number of rejections are always above 1 in 2.

The effects of adjusting for multiple-testing on the optimal-reaction hypothesis (see, e.g., Fuhrer, 2018) were qualitatively similar: all corrections except for FDR-BH tend to markedly reduce the respondent rejection rates to low levels, especially for the short-horizon forecasts.

For the second paper, the study of forecasters' perceptions of output growth persistence by Clements (2020), the effects of making an allowance for multiple testing depend on whether we consider the 'medium term' or the 'long term'. There is little effect from any of the FWER or FDR approaches on the number of rejections when we consider perceptions of medium-term growth prospects. For the 10-year forecasts, however, all the approaches (other than FDR-BH and k-FWER) reduce the proportion of rejections from just under 1 in 2 to less than 1 in 10.

Finally, in the third study of the accuracy of individual forecasters' uncertainty assessments by Clements (2014), making any adjustment for multiple testing removes the evidence for over-confidence at the longer horizons, but leaves the evidence for under-confidence at within-year horizons for inflation largely intact.

We have adopted the conservative assumption that $\pi_0 = 1$ in implementing FDR control, but in the Appendix, for selective cases, indicate how the results would change for π_0 less than one.

Overall our findings suggest the researcher would do well to consider multiple-testing adjustments when analysing individual survey expectations. However, it may not matter much whether FWER or FDR control is used given the likely numbers of tests that would be run for quarterly macro survey data.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The original forecast data is from the U.S. Survey of Professional Forecasters, as described in Section 4, where a link to the website is given.

Acknowledgements

No external funding was received for this paper.

The author gratefully acknowledges helpful comments from participants at the International Symposium of Forecasting, Charlottesville, 2023.

Appendix

Table 6 provides the details of the inference made for each individual respondent, for current-quarter consumption growth forecast efficiency (MZ) tests. (The aggregate results are given in Table 3, Part A column 1). The first column gives the respondent identifier, and the second column the sorted p-values. The liberal approach of ignoring MT issues results in rejecting for the 56% of respondents with p-values smaller than 0.05 (at the 5% level: the respondents above and including respondent id463). With Bonferroni correction (column (8)), we only reject the null for 16% of respondents. (In this column, a 1 indicates Reject. The proportion of rejections is shown in the final row). For Holm in column (9), we present the α_i values from Eq. (1), which are compared to the sorted p -values, $p_{(i)}$, and turns out to give the same rejections as Bonferroni. Column (10) records the values of α_i following (2) for Holm k-FWER control. Column (4) indicates rejections for FDR BH, from comparing the sorted p-values to the values in column (3), BY (column (5)) results in fewer p-values being deemed significant relative to BH (column (4)). Column (6) records the q-values for each test statistic. The interpretation of the q-value of 0.03739 for id 518 is that a rejection of the null for this respondent comes with a concomitant expected proportion of false positives of 4.91% (which is admissible for our assumed FDR of 5%). The next largest q-value is 0.05176 (id 557), and we do not reject, as to do so would occur with a FDR above 5%. From Eq. (11) it is apparent that setting $\hat{\pi}_0 = \frac{1}{2}$, say, will double the values in column (3). The effect of different values of π_0 can easily be determined — for $\hat{\pi}_0 = \frac{1}{2}$ the proportion of rejections increases from 0.44 for BH to 0.56, the proportion of rejections when no allowance is made for MT. The effects on other approaches such as BY can also easily be determined. Column (7) records the α_i (defined in (6)) against which the p -values are compared to control the FDP. FDP control delivers the same rejections as Holm.

Table 7 provides another example, for the 4-quarter consumption growth reaction-to-news (The aggregate results are given in Table 3, Part B, column 2). In this case, setting $\hat{\pi}_0 = \frac{1}{2}$ increases the BH rejection rate of 0.25 to 0.31, compared to the liberal strategy of 0.37.

Table 6
Illustration of the effects of Bonferroni correction and FDR control for forecast efficiency tests of $h = 0$ forecasts of Consumption.

Individual	p-value	$(i/N) \times q$	BH	BY	q-value	FDP	Bonf.	Holm α_i	Holm k-FWER α_i
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
20	0.0000	0.001	1	1	0.0000	0.0010	1	0.0010	0.0020
548	0.0000	0.002	1	1	0.0002	0.0010	1	0.0010	0.0020
484	0.0000	0.003	1	1	0.0003	0.0010	1	0.0010	0.0020
99	0.0000	0.004	1	1	0.0005	0.0011	1	0.0011	0.0021
428	0.0001	0.005	1	1	0.0006	0.0011	1	0.0011	0.0021
456	0.0002	0.006	1	1	0.0019	0.0011	1	0.0011	0.0022
426	0.0009	0.007	1	1	0.0059	0.0011	1	0.0011	0.0022
498	0.0009	0.008	1	1	0.0059	0.0012	1	0.0012	0.0023
420	0.0015	0.009	1	1	0.0083	0.0012	0	0.0012	0.0023
512	0.0017	0.010	1	1	0.0083	0.0012	0	0.0012	0.0024
433	0.0037	0.011	1	0	0.0161	0.0013	0	0.0013	0.0024
483	0.0039	0.012	1	0	0.0161	0.0013	0	0.0013	0.0025
524	0.0042	0.013	1	0	0.0161	0.0013	0	0.0013	0.0026
407	0.0056	0.014	1	0	0.0199	0.0014	0	0.0014	0.0026
526	0.0060	0.015	1	0	0.0199	0.0014	0	0.0014	0.0027
431	0.0084	0.016	1	0	0.0262	0.0014	0	0.0014	0.0028
414	0.0114	0.017	1	0	0.0301	0.0015	0	0.0015	0.0029
446	0.0114	0.018	1	0	0.0301	0.0015	0	0.0015	0.0029
439	0.0115	0.019	1	0	0.0301	0.0016	0	0.0016	0.0030
507	0.0128	0.020	1	0	0.0319	0.0031	0	0.0016	0.0031
508	0.0159	0.021	1	0	0.0375	0.0032	0	0.0017	0.0032
518	0.0165	0.022	1	0	0.0375	0.0033	0	0.0017	0.0033
557	0.0249	0.023	0	0	0.0518	0.0034	0	0.0018	0.0034
429	0.0255	0.024	0	0	0.0518	0.0036	0	0.0019	0.0036
506	0.0259	0.025	0	0	0.0518	0.0037	0	0.0019	0.0037
556	0.0284	0.026	0	0	0.0546	0.0038	0	0.0020	0.0038
405	0.0301	0.027	0	0	0.0558	0.0040	0	0.0021	0.0040
463	0.0471	0.028	0	0	0.0842	0.0042	0	0.0022	0.0042
527	0.0641	0.029	0	0	0.1086	0.0043	0	0.0023	0.0043
84	0.0654	0.030	0	0	0.1086	0.0045	0	0.0024	0.0045
535	0.0673	0.031	0	0	0.1086	0.0048	0	0.0025	0.0048
94	0.0714	0.032	0	0	0.1100	0.0050	0	0.0026	0.0050
423	0.0734	0.033	0	0	0.1100	0.0053	0	0.0028	0.0053
421	0.0748	0.034	0	0	0.1100	0.0056	0	0.0029	0.0056
510	0.0812	0.035	0	0	0.1161	0.0059	0	0.0031	0.0059
411	0.1046	0.036	0	0	0.1452	0.0063	0	0.0033	0.0063
542	0.1252	0.037	0	0	0.1692	0.0067	0	0.0036	0.0067
40	0.1437	0.038	0	0	0.1890	0.0071	0	0.0038	0.0071
422	0.1612	0.039	0	0	0.2067	0.0077	0	0.0042	0.0077
516	0.1764	0.040	0	0	0.2205	0.0115	0	0.0045	0.0083
472	0.1863	0.041	0	0	0.2271	0.0125	0	0.0050	0.0091
546	0.1959	0.042	0	0	0.2332	0.0136	0	0.0056	0.0100
555	0.2328	0.043	0	0	0.2708	0.0150	0	0.0063	0.0111
553	0.2703	0.044	0	0	0.3071	0.0167	0	0.0071	0.0125
520	0.2902	0.045	0	0	0.3203	0.0188	0	0.0083	0.0143
504	0.2947	0.046	0	0	0.3203	0.0214	0	0.0100	0.0167
404	0.5058	0.047	0	0	0.5381	0.0250	0	0.0125	0.0200
540	0.5886	0.048	0	0	0.6131	0.0300	0	0.0167	0.0250
528	0.6270	0.049	0	0	0.6398	0.0375	0	0.0250	0.0333
424	0.8193	0.050	0	0	0.8193	0.0500	0	0.0500	0.0500
	0.56		0.44	0.20	.	0.16	0.16	0.16	0.20

Table 7
Illustration of the effects of Bonferroni correction and FDR control for the reaction-to-news tests of $h = 4$ forecasts of Consumption.

Individual	p-value	$(i/N) \times q$	BH	BY	q-value	FDP	Bonf.	Holm α_i	Holm k-FWER α_i
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
20	0.0000	0.0010	1	1	0.0000	0.0010	1	0.0010	0.0020
405	0.0000	0.0020	1	1	0.0000	0.0010	1	0.0010	0.0020
99	0.0000	0.0031	1	1	0.0000	0.0011	1	0.0011	0.0021
414	0.0000	0.0041	1	1	0.0003	0.0011	1	0.0011	0.0021
498	0.0001	0.0051	1	1	0.0010	0.0011	1	0.0011	0.0022
404	0.0003	0.0061	1	1	0.0025	0.0011	1	0.0011	0.0022
527	0.0015	0.0071	1	1	0.0101	0.0012	0	0.0012	0.0023
535	0.0016	0.0082	1	1	0.0101	0.0012	0	0.0012	0.0023
439	0.0022	0.0092	1	0	0.0117	0.0012	0	0.0012	0.0024
94	0.0042	0.0102	1	0	0.0204	0.0013	0	0.0013	0.0024
507	0.0046	0.0112	1	0	0.0204	0.0013	0	0.0013	0.0025
557	0.0118	0.0122	1	0	0.0483	0.0013	0	0.0013	0.0026
548	0.0159	0.0133	0	0	0.0600	0.0014	0	0.0014	0.0026
426	0.0258	0.0143	0	0	0.0887	0.0014	0	0.0014	0.0027
431	0.0271	0.0153	0	0	0.0887	0.0014	0	0.0014	0.0028
506	0.0368	0.0163	0	0	0.1080	0.0015	0	0.0015	0.0029
429	0.0375	0.0173	0	0	0.1080	0.0015	0	0.0015	0.0029
504	0.0464	0.0184	0	0	0.1263	0.0016	0	0.0016	0.0030
456	0.0553	0.0194	0	0	0.1426	0.0016	0	0.0016	0.0031
84	0.0599	0.0204	0	0	0.1469	0.0032	0	0.0017	0.0032
508	0.0710	0.0214	0	0	0.1609	0.0033	0	0.0017	0.0033
421	0.0722	0.0224	0	0	0.1609	0.0034	0	0.0018	0.0034
483	0.0883	0.0235	0	0	0.1880	0.0036	0	0.0019	0.0036
433	0.1087	0.0245	0	0	0.2219	0.0037	0	0.0019	0.0037
484	0.1166	0.0255	0	0	0.2285	0.0038	0	0.0020	0.0038
555	0.1354	0.0265	0	0	0.2551	0.0040	0	0.0021	0.0040
542	0.1494	0.0276	0	0	0.2712	0.0042	0	0.0022	0.0042
446	0.1749	0.0286	0	0	0.3061	0.0043	0	0.0023	0.0043
40	0.1845	0.0296	0	0	0.3118	0.0045	0	0.0024	0.0045
520	0.2310	0.0306	0	0	0.3773	0.0048	0	0.0025	0.0048
540	0.2397	0.0316	0	0	0.3789	0.0050	0	0.0026	0.0050
528	0.2633	0.0327	0	0	0.4032	0.0053	0	0.0028	0.0053
411	0.2974	0.0337	0	0	0.4299	0.0056	0	0.0029	0.0056
407	0.2983	0.0347	0	0	0.4299	0.0059	0	0.0031	0.0059
510	0.4027	0.0357	0	0	0.5540	0.0063	0	0.0033	0.0063
526	0.4070	0.0367	0	0	0.5540	0.0067	0	0.0036	0.0067
518	0.4309	0.0378	0	0	0.5706	0.0071	0	0.0038	0.0071
424	0.5142	0.0388	0	0	0.6631	0.0077	0	0.0042	0.0077
422	0.5589	0.0398	0	0	0.7022	0.0083	0	0.0045	0.0083
546	0.6012	0.0408	0	0	0.7364	0.0125	0	0.0050	0.0091
428	0.6274	0.0418	0	0	0.7498	0.0136	0	0.0056	0.0100
524	0.6840	0.0429	0	0	0.7980	0.0150	0	0.0063	0.0111
420	0.8290	0.0439	0	0	0.9446	0.0167	0	0.0071	0.0125
516	0.8747	0.0449	0	0	0.9741	0.0188	0	0.0083	0.0143
512	0.9245	0.0459	0	0	0.9764	0.0214	0	0.0100	0.0167
553	0.9354	0.0469	0	0	0.9764	0.0250	0	0.0125	0.0200
556	0.9447	0.0480	0	0	0.9764	0.0300	0	0.0167	0.0250
472	0.9565	0.0490	0	0	0.9764	0.0375	0	0.0250	0.0333
463	0.9952	0.0500	0	0	0.9952	0.0500	0	0.0500	0.0500
	0.37	.	0.25	0.16	.	0.12	0.12	0.12	0.18

References

Barras, L., Scaillet, O., Wermers, R., 2010. False discoveries in mutual fund performance: Measuring luck in estimated alphas. J. Finance 65 (1), 179–216.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B Stat. Methodol. 57 (1), 289–300.

Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. Ann. Statist. 29 (4), 1165–1188.

Bluedorn, J.C., Leigh, D., 2018. Is the Cycle the Trend? Evidence From the Views of International Forecasters. Working Paper 18/163, International Monetary Fund.

Bonferroni, C.E., 1936. Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni Istituto Super. Sci. Econ. Commer. Firenze 8, 3–62.

Bordalo, P., Gennaioli, N., Ma, Y., Shleifer, A., 2020. Overreaction in macroeconomic expectations. Amer. Econ. Rev. 110 (9), 2748–2782.

Broer, T., Kohlhas, A., 2021. Forecaster (Mis-)Behavior. CEPR Discussion Papers 12898/2018, C.E.P.R. Discussion Papers.

Cheng, L., Sheng, X.S., 2017. Combination of "combinations of p values". Empir. Econ. 53 (1), 329–350.

Chordia, T., Goyal, A., Saretto, A., 2020. Anomalies and false rejections. Rev. Financ. Stud. 33 (5), 2134–2179.

Clements, M.P., 2009. Internal consistency of survey respondents' forecasts: Evidence based on the survey of professional forecasters. In: Castle, J.L., Shephard, N. (Eds.), The Methodology and Practice of Econometrics. A Festschrift in Honour of David F. Hendry. Oxford University Press, Oxford, pp. 206–226 (Chapter 8).

Clements, M.P., 2014. Forecast uncertainty - ex ante and ex post: US inflation and output growth. J. Bus. Econom. Statist. 32 (2), 206–216. http://dx.doi.org/10.1080/07350015.2013.859618.
Clements, M.P., 2020. Individual forecaster perceptions of the persistence of shocks to GDP. J. Appl. Econometrics 37, 640–656.
Clements, M.P., 2022. Forecaster efficiency, accuracy and disagreement: Evidence using individual-level survey data. J. Money Credit Bank. 54, 537–567, Nos. 2-3.
Clements, M.P., 2024. Do professional forecasters believe in the Phillips curve? Int. J. Forecast. 40 (3), 1238–1254.
Clements, M.P., Rich, R., Tracy, J., 2023. Surveys of professionals, chapter 3. In: Ruediger Bachmann, W.K. (Ed.), Handbook of Economic Expectations. Academic Press, Elsevier, pp. 71–106.
Coibion, O., Gorodnichenko, Y., 2012. What can survey forecasts tell us about information rigidities? J. Polit. Econ. 120 (1), 116–159.
Coibion, O., Gorodnichenko, Y., 2015. Information rigidity and the expectations formation process: A simple framework and new facts. Amer. Econ. Rev. 105 (8), 2644–2678.
Croushore, D., 1993. Introducing: The survey of professional forecasters. Fed. Reserve Bank Phila. Bus. Rev. Novemb. 3–15.
Croushore, D., Stark, T., 2001. A real-time data set for macroeconomists. J. Econometrics 105 (1), 111–130.
Efron, B., Hastie, T., 2016. Computer age statistical inference. In: Institute of Mathematical Statistics Monographs, Cambridge University Press.
Engelberg, J., Manski, C.F., Williams, J., 2009. Comparing the point predictions and subjective probability distributions of professional forecasters. J. Bus. Econom. Statist. 27 (1), 30–41.
Fuhrer, J.C., 2018. Intrinsic Expectations Persistence: Evidence from Professional and Household Survey Expectations. Working Papers 18-9, FRB of Boston.
Glickman, M.E., Rao, S.R., Schultz, M.R., 2014. False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. J. Clin. Epidemiol. 67, 850–857.
Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 65–70.
Jain, M., 2019. Perceived inflation persistence. J. Bus. Econom. Statist. 37 (1), 110–120.
Knüppel, M., Schultefrankenfeld, G., 2019. Assessing the uncertainty in central banks' inflation outlooks. Int. J. Forecast. 35, 1748–1769.
Krane, S.D., 2011. Professional forecasters' view of permanent and transitory shocks to GDP. Am. Econ. J.: Macroecon. 3 (1), 184–211.
Lehmann, E.L., Romano, J.P., 2005. Generalizations of the familywise error rate. Ann. Statist. 33 (3), 1138–1154.
Mankiw, N.G., Reis, R., 2002. Sticky information versus sticky prices: A proposal to replace the New Keynesian Phillips curve. Q. J. Econ. 117, 1295–1328.
Mankiw, N.G., Reis, R., Wolfers, J., 2003. Disagreement about inflation expectations. mimeo, National Bureau of Economic Research, Cambridge MA.
Mayo, D.G., 2018. Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars. Cambridge University Press.
Mincer, J., Zarnowitz, V., 1969. The evaluation of economic forecasts. In: Mincer, A. (Ed.), Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance. National Bureau of Economic Research, New York, pp. 3–46.
Romano, J.P., Wolf, M., 2005. Stepwise multiple testing as formalized data snooping. Econometrica 73 (4), 1237–1282.
Simes, R.J., 1986. An improved Bonferroni procedure for multiple tests of significance. Biometrika 73 (3), 751–754.
Sims, C.A., 2003. Implications of rational inattention. J. Monetary Econ. 50, 665–690.
Storey, J.D., Tibshirani, R., 2003. Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. 100 (16), 9440–9445.
White, H., 2000. A reality check for data snooping. Econometrica 68, 1097–1126.
Woodford, M., 2002. Imperfect common knowledge and the effects of monetary policy. In: Aghion, P., Frydman, R., Stiglitz, J., Woodford, M. (Eds.), Knowledge, Information, and Expectations in Modern Macroeconomics: in Honor of Edmund Phelps. Princeton University Press, pp. 25–58.