Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

The Sum and Its Parts: Judgmental Hierarchical Forecasting

Mirko Kremer, Enno Siemsen, Douglas J. Thomas

To cite this article:
Mirko Kremer, Enno Siemsen, Douglas J. Thomas (2016) The Sum and Its Parts: Judgmental Hierarchical Forecasting. Management Science 62(9):2745-2764. https://doi.org/10.1287/mnsc.2015.2259

MANAGEMENT SCIENCE
Vol. 62, No. 9, September 2016, pp. 2745–2764
ISSN 0025-1909 (print) | ISSN 1526-5501 (online)

http://dx.doi.org/10.1287/mnsc.2015.2259

The Sum and Its Parts: Judgmental Hierarchical Forecasting

Mirko Kremer
Management Department, Frankfurt School of Finance and Management, 60314 Frankfurt am Main, Germany, m.kremer@fs.de

Enno Siemsen
Carlson School of Management, University of Minnesota, Minneapolis, Minnesota 55455; and Wisconsin School of Business, University of Wisconsin–Madison, Madison, Wisconsin 53706, esiemsen@wisc.edu

Douglas J. Thomas
Smeal College of Business, Penn State University, University Park, Pennsylvania 16802, dthomas@psu.edu

Firms require demand forecasts at different levels of aggregation to support a variety of resource allocation decisions. For example, a retailer needs store-level forecasts to manage inventory at the store, but also requires a regionally aggregated forecast for managing inventory at a distribution center. In generating an aggregate forecast, a firm can choose to make the forecast *directly* based on the aggregated data or *indirectly* by summing lower-level forecasts (i.e., bottom up). Our study investigates the relative performance of such hierarchical forecasting processes through a behavioral lens. We identify two judgment biases that affect the relative performance of direct and indirect forecasting approaches: a propensity for random judgment errors and a failure to benefit from the informational value that is embedded in the correlation structure between lower-level demands. Based on these biases, we characterize demand environments where one hierarchical process results in more accurate forecasts than the other.

Keywords: judgmental forecasting; nonstationary demand; covariation detection; behavioral operations; bottom-up forecasting; random judgment error
History: Received August 10, 2012; accepted May 7, 2015, by Martin Lariviere, operations management. Published online in *Articles in Advance* December 18, 2015.

1. Introduction

Operational decision making requires accurate demand forecasts at different product levels (i.e., stock-keeping unit versus product family) as well as at different market levels (i.e., regional versus global). Consider a distribution center storing inventory that is used to replenish multiple retail stores. Store-level forecasts are needed to plan store operations, but an aggregate (regional) forecast is required for the inventory decisions at the distribution center. It is not immediately clear whether these aggregate forecasts should be made *directly* on the aggregated data or *indirectly* by summing up lower-level forecasts. To shed light on this issue, our study compares the performance of direct and indirect forecasting approaches.

Empirical comparisons of direct and indirect forecasting approaches are scarce and mostly focused on the statistical aspects of forecasting (Armstrong 1985). The consensus in this literature is a preference for bottom-up forecasting (Dangerfield and Morris 1992, Armstrong 2001, Allen and Fildes 2001), or to forecast at both levels and reconcile them using regression (Hyndman et al. 2011). Yet forecasting in practice is not a purely statistical task. It is (Fildes et al. 2009, Fildes and Petropoulos 2015), and arguably should be (Blattberg and Hoch 1990), influenced by the judgment of the forecaster. The judgmental forecasting literature is vast (for a review, see Lawrence et al. 2006), but is focused on behavioral anomalies in univariate settings and remains silent on the issue of aggregation in multivariate and hierarchical contexts. Because judgmental forecasting research does not study hierarchical forecasting, and hierarchical forecasting research does not study judgmental forecasting, the existing literature provides little guidance on whether judgmental forecasting becomes more accurate under direct or indirect forecasting approaches.

Are judgmental forecasts more accurate at the top level if they are made directly on aggregated data, or if they are made indirectly by summing up lower-level forecasts? And what are the demand characteristics that influence the relative performance of these two alternative forecasting processes? To address these research questions, we present the results from two studies based on a series of behavioral experiments conducted in a controlled laboratory environment. Subjects in our studies (undergraduate and

graduate students, as well as forecasting professionals) make sequential one-period-ahead forecasts at different levels of aggregation. Our research contains two important distinctions related to the demand environment. First, in contrast to the simple stationary demand environments typically studied in the behavioral operations literature (e.g., Schweitzer and Cachon 2000, Özer et al. 2011), we study forecasting in nonstationary demand environments, which may substantially change behavior (e.g., Kremer et al. 2011) and prescriptions (e.g., Graves 1999). Second, we systematically vary the nature of correlation between lower-level demand series. This allows us to capture several realistic environments that vary in terms of substitutability and complementarity between products.

Across the demand environments we study, *statistical* forecasts yield identical performance for top-direct and indirect bottom-up approaches to forecasting top-level demand. However, our data show that the accuracy of *judgmental* top-level forecasts is generally not equivalent for direct and indirect forecasting processes. The relative performance of one process compared to the other systematically depends on the correlation structure of lower-level demands and how these demand characteristics modulate the effects of two fundamental judgment biases.[1] The first bias relates to a propensity of forecasters to make random judgment errors, where the magnitude of the error is affected by the demand environment. Depending on the correlation structure of lower-level demands, the random judgment error in a bottom-up forecast (the aggregation of lower-level judgment errors) can be more or less detrimental to forecast accuracy than the judgment error in a forecast generated directly at the top level. The second bias stems from an inability of forecasters to detect and exploit the correlation between lower-level demands. In environments where lower-level demand correlations theoretically offer valuable information (which forecasters fail to capture in the bottom-up process), this judgment bias favors forecasts generated directly at the top level. Together, our studies identify demand environments where we expect one process to result in lower forecast error than the other. This allows us to gain insights that go beyond what previous (mostly statistical) research on hierarchical forecasting has suggested. Importantly, the correlation structure between item-level demands matters, because it drives relative performance of direct versus indirect approaches to forecasting top-level demand in a systematic fashion.

[1] Throughout this paper, we use *bias* to refer to *judgment* bias, as opposed to a measure of consistent over- or underforecasting commonly used in the forecasting literature.

This paper proceeds as follows. Section 2 provides an overview of the demand environment we use in our study and the normative benchmark in that environment. Sections 3 and 4 present our experimental studies: Study 1 focuses on the effect of aggregating random judgment errors, and Study 2 focuses on the aggregation of biases regarding the reaction to observed forecast errors. We discuss the managerial implications of our results in detail in §5, and conclude in §6.

2. Forecasting Environment

In this section, we describe the forecasting environment we use in our studies and provide necessary definitions. Let F_t denote a forecast made for period t, D_t realized demand in period t, and $E_t = D_t - F_t$ the corresponding forecast error. Throughout, we use mean absolute forecast error, defined as $MAE_T = T^{-1}\sum_{t=1}^{T}|E_t|$ for some horizon T, as our forecast performance metric.

Consider $i = 1..M$ items with period t demands $\mathbf{D}_t = \{D_{1,t}, \ldots, D_{M,t}\}$. Forecasters have no additional information on future demands beyond what is contained either in the lower-level multivariate time series $\{\mathbf{D}_t, \mathbf{D}_{t-1}, \mathbf{D}_{t-2}, \ldots\}$ or in the top-level series, where top-level demand in period t is simply the sum of all lower-level demands, $\ddot{D}_t = \sum_{i=1}^{M} D_{i,t}$. We will use $\ddot{D}_t$ and $\ddot{F}_t$ to indicate the aggregated demand series and a forecast for that series, respectively. Our research is concerned with two processes for forecasting top-level demand $\ddot{D}_t$. Top-level forecasts can be made directly on top-level series. We denote these forecasts, prepared for period t, by $\ddot{F}_t^{DT}$ (*DT* for *direct-top*). Alternatively, in a *bottom-up* (*BU*) process, item-level forecasts $F_{i,t}$ are made based on item-level demand information and then summed up to yield a top-level forecast $\ddot{F}_t^{BU} = \sum_{i=1}^{M} F_{i,t}$. The objective of our research is an assessment of how judgment affects the relative accuracy of these two forecasting processes, and how the behavioral advantage of one process over the other is moderated by the underlying demand conditions.

2.1. Demand Process: Correlation and Aggregation

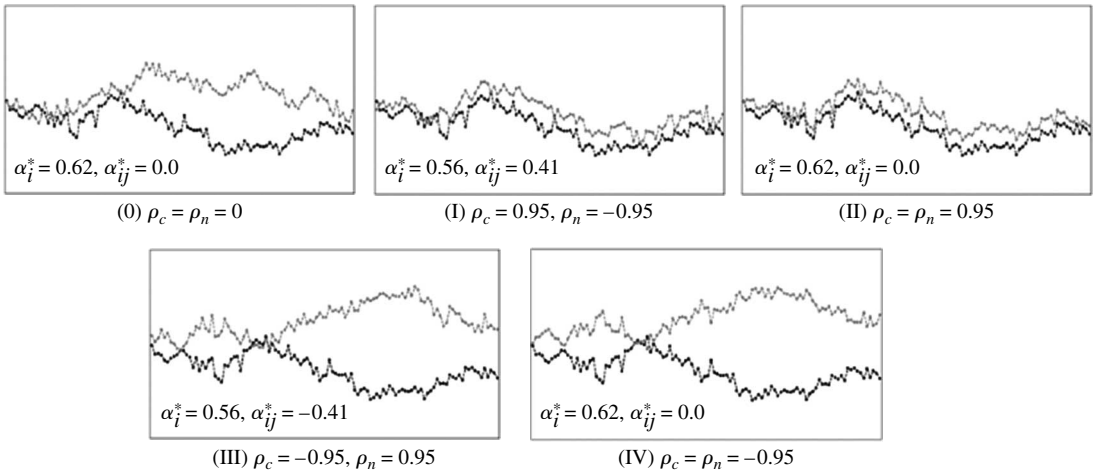
Item-level demand follows a nonstationary process:

$$\mathbf{D}_t = \boldsymbol{\mu}_t + \boldsymbol{\eta}_t, \tag{1a}$$

$$\boldsymbol{\mu}_t = \boldsymbol{\mu}_{t-1} + \boldsymbol{\varsigma}_t, \tag{1b}$$

where $\boldsymbol{\eta}_t \sim \text{Normal}(\mathbf{0}, \mathbf{N})$ and $\boldsymbol{\varsigma}_t \sim \text{Normal}(\mathbf{0}, \mathbf{C})$ are normally distributed, serially independent disturbances with zero mean and covariance matrices $\mathbf{N}$ and $\mathbf{C}$, respectively. The vector $\boldsymbol{\mu}_t$ represents the unobservable levels of the lower-level series. The forecasting task in this environment essentially requires estimating the unobserved level $\boldsymbol{\mu}_t$. Each time series contains two random components: temporary shocks

Figure 1 **Example Demand Paths**



(through $\boldsymbol{\eta}_t$, termed *noise* throughout) and permanent shocks (through $\boldsymbol{\varsigma}_t$, termed *change* throughout). To simplify our exposition, we consider only $M = 2$ symmetric items in our two studies, each with a standard deviation of change c and a standard deviation of noise n. Change random variables for the two items are correlated with correlation coefficient ρ_c; noise random variables are correlated with correlation coefficient ρ_n. The demand process in (1a)–(1b) has a number of appealing properties: it provides a flexible way for modeling substitution patterns, an intuitive aggregation mechanism, and a simple normative forecasting benchmark.

Through the correlation coefficients ρ_c and ρ_n, we can represent a variety of business situations. For example, one might expect demands for different items in the same product family in the same location to be subject to the same persistent effects (positively correlated change), such as general economic conditions or brand-level advertising campaigns. Similarly, items may be complementary, such that increases in demand for one item naturally lead to increases in demand for the other. Alternatively, consumers may substitute among goods in the same product category. Such substitution effects could be transient (negatively correlated noise), where a consumer chooses a particular item one day without a lasting effect, or permanent (negatively correlated change), where one item becomes preferred at the expense of another in the long run. Positively correlated noise may occur, for example, when a transient shock, such as a weather event, affects short-term sales of several items in a similar manner. Thus, the demand process defined in Equations (1a) and (1b) enables us to capture a rich set of complementarity and substitution effects between items. In Figure 1, we provide example data sets for different values of ρ_c and ρ_n. Positive (negative) values of ρ_n are associated with "spikes" in the time series that go in the same (opposite) directions across time series. Positive (negative) values of ρ_c imply that the time series move together (apart) in the long run.

2.2. Bottom-Up Forecasting

Multivariate exponential smoothing provides the optimal forecasting model for the demand process described in Equations (1a) and (1b) (Jones 1966). For two items i, j,

$$F_{i,t+1} = F_{i,t} + \alpha_{ii}E_{i,t} + \alpha_{ij}E_{j,t}, \tag{2a}$$

$$F_{j,t+1} = F_{j,t} + \alpha_{jj}E_{j,t} + \alpha_{ji}E_{i,t}, \tag{2b}$$

where the forecast adjustment for series i (from $F_{i,t}$ to $F_{i,t+1}$) is a reaction to the observed focal forecast error ($E_{i,t}$) as well as the distal forecast error ($E_{j,t}$). The strength of these reactions depends on the smoothing weights ($\alpha_{ii}, \alpha_{ij}, \alpha_{ji}, \alpha_{jj}$). To build some intuition for the link between the demand and forecasting processes, consider first the case of two independent demand series. Under independence, Equations (2a) and (2b) reduce to single exponential smoothing, i.e., $F_{t+1} = F_t + \alpha E_t$, where a previous forecast F_t is adjusted toward the current demand observation D_t. The direction and magnitude of this forecast adjustment depends on the forecast error $E_t = D_t - F_t$, as well as the smoothing weight α, which corresponds to a reaction parameter that determines how strongly a forecasting process reacts to observed errors. A key concept to understand how strong this reaction should be is the change-to-noise ratio:

$$W = \frac{c^2}{n^2}. \tag{3}$$

This ratio captures the "weight of evidence" inherent in an observed demand variation and is a measure of how much of an observed variation in the series is

due to permanent versus temporary shocks. Furthermore, W intuitively maps into the forecasting mechanism, as the optimal smoothing constant[2] for single exponential smoothing is (Harrison 1967)

$$\alpha^* = \frac{2}{1+\sqrt{1+4W^{-1}}}. \tag{4}$$

Equation (4) captures the intuition behind the key challenge of how to optimally separate true change in the demand level (through c) from random noise (through n). Previous forecasts should be adjusted more heavily (i.e., large α^*) in environments where variations in demand are indicative of level changes (large W), whereas variations in demand should be mostly discarded (i.e., small α^*) when they are mostly due to noise (small W).

The multivariate demand process in Equations (1a) and (1b) gives rise to an $M \times M$ matrix of optimal smoothing weights, which include focal-error-response coefficients α_{ii} for the focal error $E_{i,t}$ as well as distal-error-response coefficients α_{ij} for the distal error $E_{j,t}$ (Equations (2a) and (2b)). Although there are no simple expressions for optimal multivariate smoothing constants analogous to Equation (4) for the univariate setting,[3] we can develop an intuition for the link between optimal forecasting and the demand process for the two-item setting we consider in this study. To see why the forecast errors from a distal series (series j, in this case) can be helpful in disentangling change from noise for a focal series (series i, in this case), consider the situation where change draws are positively correlated, but noise draws are independent ($\rho_c > 0$, $\rho_n = 0$). In this case, observing forecast errors $(E_{i,t}, E_{j,t})$ with the same sign in the two related series provides stronger evidence that the observed demand variation is indicative of a persistent change in the level of the time series, whereas errors with opposite signs suggest more of the forecast error should be attributed to noise. Thus, the optimal exponential smoothing forecast places nonnegative weights on forecast errors for both items (Figure 1(I)). In fact, this benefit is strengthened if the noise draws are negatively correlated; observing errors with the same sign provides even stronger confirmation of underlying change. Analogous intuition follows if the change terms are negatively correlated and noise terms are positively correlated; the weight placed on the distal forecast error under such conditions is negative (Figure 1(III)). The optimal forecasting process reduces to univariate single exponential smoothing ($\alpha^*_{ij} = \alpha^*_{ji} = 0$) when the data series are independent ($\rho_c = \rho_r = 0$; see Figure 1(0)), or when the change and noise covariance matrices are proportional (i.e., if $\mathbf{C} = q\mathbf{N}$; Enns et al. 1982, Harvey 1986). Figures 1(II) and 1(IV) illustrate two such demand environments, where the distal time series provides no information for focal forecasts ($\alpha^*_{ij} = 0$) even though both series are strongly positively (negatively) correlated, i.e., $\rho_c = \rho_n = 0.95$ ($= -0.95$).

In the context of a bottom-up forecasting process with two symmetric items, and letting $\ddot{E}^{BU}_t = E_{i,t} + E_{j,t}$, the forecasts from Equations (2a) and (2b) provide the *implied* top-level forecast, $\ddot{F}^{BU}_{t+1} = F_{i,t+1} + F_{j,t+1} = \ddot{F}^{BU}_t + (\alpha^*_{ii} + \alpha^*_{ij})\ddot{E}^{BU}_t$.

2.3. Direct-Top Forecasting

The *direct-top* forecast is defined by $\ddot{F}^{DT}_{t+1} = \ddot{F}^{DT}_t + \ddot{\alpha}^* \ddot{E}^{DT}_t$, and the optimal smoothing constant $\ddot{\alpha}^*$ depends on the top-level change-to-noise ratio $\ddot{W}$ through Equation (4). For our purposes, it is important to understand the structural properties of top-level demand $\ddot{D}_t$ as a function of the properties of the lower-level demand process $\mathbf{D_t}$. Although aggregation preserves the structural properties of the time series (Lütkepohl 2007), the aggregation of two symmetric time series yields a change-to-noise ratio of the top-level series,

$$\ddot{W} = \frac{(1+\rho_c)}{(1+\rho_n)} W.^{4} \tag{5}$$

Equation (5) describes how the top-level demand series carries more (less) weight of evidence than a single item-level series (Equation (3)), depending on the correlation structure between the item-level series. Intuitively, the change-to-noise ratio is large at the top level compared to the lower level when ρ_c is large and/or ρ_n is small (or negative). Importantly, there is a simple relationship between optimal direct-top and bottom-up forecasting (Lütkepohl 2007),

$$\ddot{\alpha}^* = \alpha^*_{ii} + \alpha^*_{ij}. \tag{6}$$

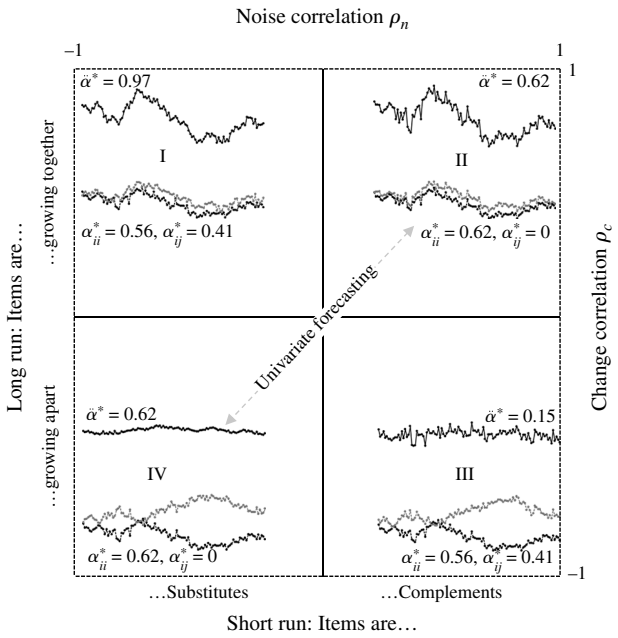
Figure 2 illustrates this relationship. When $\rho_c > \rho_n$ (quadrant I), the change-to-noise ratio is greater at the top level than at the lower level ($\ddot{W} > W$), which requires a larger smoothing parameter $\ddot{\alpha}^*$ at the top level than the focal α^*_{ii} at the lower level. When $\rho_c < \rho_n$

[2] This smoothing constant is optimal in the sense that it produces unbiased forecast errors with minimum variance. Given that forecast errors are normally distributed in our setting, this value of the smoothing constant also minimizes mean absolute error.

[3] Finding optimal smoothing constants requires recursive solution of a system of equations, even in the $M = 2$ product case. For the general case, the optimal forecasting process is given by $\mathbf{F}_{t+1} = \mathbf{F}_t + \mathbf{A}^*(\mathbf{D}_t - \mathbf{F}_t)$, where $\mathbf{F}_t$ and $\mathbf{D}_t$ are $M \times 1$ vectors, and $\mathbf{A}^*$ is an $M \times M$ matrix. The optimal smoothing matrix can be numerically determined as $\mathbf{A}^* = I - N(S^* + N)^{-1}$, where S^* is the (unique) solution to $S = N(S+N)^{-1}S + C$ (see Jones 1966). For easy reference, Figure 1 contains the optimal smoothing parameters for the specific demand environments we implement in our behavioral experiments.

[4] Variances of change and noise for the resulting top-level series are given by $\ddot{c}^2 = 2c^2(1+\rho_c)$ and $\ddot{n}^2 = 2n^2(1+\rho_n)$.

Figure 2 Aggregation of Demand and Forecasting Process



(quadrant III), the change-to-noise ratio is smaller at the top level than at the lower level ($\ddot{W} < W$), which requires a smaller smoothing parameter $\ddot{\alpha}^*$ at the top level than the focal α^*_{ii} at the lower level. (If the change and noise components of the lower-level series have equal correlations (i.e., $\rho_c = \rho_n$), then $\ddot{W} = W$, implying that $\ddot{\alpha}^* = \alpha^*_{ii}$, and $\alpha^*_{ij} = 0$ (quadrants II and IV).

In summary, our forecasting environment is characterized by the change and noise correlations among the lower-level series, creating a two-by-two matrix that describes how demand for lower-level items is interrelated. There is a diagonal across this matrix that shows the points where the optimal lower-level forecasting process is univariate (single exponential smoothing), and interrelations among items can therefore be safely ignored in the bottom-up forecasting process. The farther a set of items is away from this diagonal, the more important multivariate forecasting becomes. Note that the four different quadrants of this matrix, represented by the Roman numerals I–IV in Figure 2, essentially become experimental conditions in our study; we use these numerals throughout our paper to refer to these quadrants.

2.4. Judgmental Hierarchical Forecasting

The key implication from Equation (6) is that the optimal bottom-up and direct-top forecasting processes yield identical top-level forecasts, resulting in identical forecasting performance (e.g., measured by mean absolute errors). However, we predict relative performance differences based on judgment biases and the particular way these biases aggregate. To conceptualize the ensuing arguments, suppose that forecasters follow the mechanics of an exponential smoothing process when preparing their forecasts,[5] but do so imperfectly. In particular, we posit that the human reaction to observed forecast errors systematically deviates from normative predictions (captured through α^*_{ii}, α^*_{ij}, and $\ddot{\alpha}^*$). Furthermore, we consider the role of random judgment error as a source of forecasting performance loss (Bowman 1963). The notion that decision makers, when faced with the same decision context, randomly deviate from a decision rule in a form of "trembling hands" is a cornerstone in quantal response models (Su 2008, Allon et al. 2013), as well as in the wisdom of crowds literature (Larrick and Soll 2006). To formally capture a forecaster's tendency to randomly deviate from her exponential smoothing forecasts, we introduce random judgment error terms, $\varepsilon_{i,t} \sim N(0, \sigma)$ and $\ddot{\varepsilon}^{DT}_t \sim N(0, \ddot{\sigma}^{DT})$.

2.4.1. Bottom-Up. Forecasts for two lower-level series (i and j) are described by

$$F_{i,t+1} = F_{i,t} + \hat{\alpha}_{ii}E_{i,t} + \hat{\alpha}_{ij}E_{j,t} + \varepsilon_{i,t}, \tag{7a}$$

$$F_{j,t+1} = F_{j,t} + \hat{\alpha}_{ii}E_{j,t} + \hat{\alpha}_{ij}E_{i,t} + \varepsilon_{j,t}. \tag{7b}$$

Equations (7a) and (7b) assume that forecasters use identical parameters $\hat{\alpha}_{ii}$ and $\hat{\alpha}_{ij}$ across both lower-level time series, which is a reasonable assumption for the symmetric lower-level series we consider in our study. We can then write the *implied* bottom-up forecasts for the top-level series as

$$\ddot{F}^{BU}_{t+1} = F_{i,t+1} + F_{j,t+1} = \ddot{F}^{BU}_t + (\hat{\alpha}_{ii} + \hat{\alpha}_{ij})\ddot{E}^{BU}_t + \varepsilon_{i,t} + \varepsilon_{j,t}. \tag{7c}$$

2.4.2. Direct-Top. Similarly, direct-top forecasts are described by

$$\ddot{F}^{DT}_{t+1} = \ddot{F}^{DT}_{t+1} + \hat{\alpha}\ddot{E}^{DT}_t + \ddot{\varepsilon}^{DT}_t. \tag{7d}$$

Our two studies are designed to tease apart behavioral factors that drive the relative performance of judgmental bottom-up forecasting (Equations (7a)–(7c)) versus direct-top forecasting (Equation (7d)). Study 1 focuses on the aggregation of random judgment errors. Study 2 focuses on the aggregation of biases regarding the reaction to observed forecast errors.

[5] Previous research shows that exponential smoothing is a reasonable description of actual forecasting behavior (Andreassen and Kraus 1990, Lawrence and O'Connor 1992). Furthermore, in the econometric analysis of our data, we will relax this assumption, and allow for various other behaviors, such as the perception and projection of illusionary trends.

3. Study 1: Random Judgment Error and the Aggregation of Forecasts

To focus on the implications of random judgment errors (captured by $\varepsilon_{i,t}$ and $\ddot{\varepsilon}_t^{DT}$ in Equations (7a)–(7d)), Study 1 concentrates on demand conditions that require univariate item-level forecasting (i.e., $\alpha_{ij}^* = 0$). This allows us to control for possible performance differences between direct and indirect forecasting approaches that may arise due to the complexity of multivariate forecasting on the item level (which requires $\alpha_{ij}^* \neq 0$).

3.1. Hypothesis Development

The premise for our main hypothesis is that the magnitude of random judgment error in forecasts (measured by σ_i and $\ddot{\sigma}^{DT}$) is directly related to the uncertainty inherent in the time series on which the forecasts are based. To measure the inherent *uncertainty of a time series*, we use the standard deviation of forecast errors obtained from optimal single exponential smoothing (e.g., Harrison 1967):

$$\Sigma_i = n/\sqrt{1-\alpha_i^*}, \tag{8a}$$

$$\ddot{\Sigma} = \ddot{n}/\sqrt{1-\ddot{\alpha}^*}. \tag{8b}$$

Because Study 1 focuses on demand conditions where univariate exponential smoothing is optimal ($\alpha_{ij}^* = 0$), the optimal reaction to forecast errors is the same on both levels of aggregation ($\alpha_i^* = \ddot{\alpha}^*$). As a result, the difference between item-level and top-level demand uncertainty in Equations (8a)–(8b) is characterized entirely by the relative level of noise (n versus $\ddot{n}$).

An important relationship for our study is that the magnitude of random judgment error inherent in a judgmental forecast increases in the underlying uncertainty of the time series (Harvey 1995, Harvey et al. 1997). This effect may be driven by a desire of decision makers for their forecast series to resemble the actual time series (Harvey 1995). A key aspect of normative forecasts (such as those resulting from exponential smoothing) is that they filter out noise, and thus appear much less variable than the underlying demand series. This creates a visual disconnect between the forecast series and the demand series. A forecaster's judgment will try to counter this disconnect by making the forecast series appear more like the underlying demand series. A study that has proposed a similar relationship is Gaur et al. (2007), who establish that the uncertainty of demand actually relates to the dispersion of opinions among experts. Lee and Siemsen (2015) also establish that, in a newsvendor decision-making framework, the standard deviation of random judgment error in point forecasts increases in the standard deviation of the underlying demand series. To simplify the following exposition, let us assume for now a linear relationship between random judgment error and the uncertainty of a time series, i.e., $\sigma_i = m\Sigma_i$.

That such a relationship between demand uncertainty and random judgment error exists is perhaps not surprising, but the implications of this relationship for hierarchical forecasting are profound. On the one hand, there is a single random judgment error in a direct-top forecast ($\ddot{\varepsilon}_t^{DT}$ in Equation (7d)), with standard deviation $\ddot{\sigma}^{DT} = m\ddot{\Sigma}$. On the other hand, the random judgment error in a bottom-up forecast is the aggregation of item-level judgment errors ($\varepsilon_{i,t} + \varepsilon_{j,t}$ in Equation (7c)). To simplify the exposition, assume for now that item-level judgment errors are uncorrelated, i.e., $\varepsilon_{i,t} \sim N(0, m\Sigma_i)$ and $\mathrm{Cov}(\varepsilon_{i,t}, \varepsilon_{j,t}) = 0$, resulting in the standard deviations of random judgment error for bottom-up forecasts $\ddot{\sigma}^{BU} = \sqrt{2}m\Sigma_i$. To understand where one forecasting process may be favored in terms of random judgment error, we calculate the ratio of such errors as[6]

$$\frac{\ddot{\sigma}^{DT}}{\ddot{\sigma}^{BU}} = \frac{m\ddot{\Sigma}}{\sqrt{2}m\Sigma_i} = \sqrt{\frac{(1+\rho_n)(1-\alpha_i^*)}{(1-\ddot{\alpha}^*)}}, \tag{9}$$

where a value greater than 1 would suggest that bottom-up forecasting produces smaller random judgment errors than direct-top forecasting. Since the optimal smoothing constant for the two processes is identical when $\alpha_{ij}^* = 0$ (or, equivalently, $\ddot{\alpha}^* = \alpha_i^*$), as is the case for the demand conditions considered in Study 1, the ratio in Equation (9) simplifies to $\ddot{\sigma}^{DT}/\ddot{\sigma}^{BU} = \sqrt{(1+\rho_n)}$. This ratio suggests that bottom-up forecasting is likely to result in smaller random judgment errors than direct-top forecasting when both change and noise components across items are positively correlated (quadrant II in Figure 2). Intuitively, this effect is due to the high inherent uncertainty in the top-level series. Conversely, in quadrant IV of Figure 2, the top-level series would have very little inherent uncertainty, so we would expect direct-top random judgment errors to be small in this quadrant. We hypothesize the following:

Hypothesis 1 (Random Judgment Error). *If optimal lower-level forecasting is univariate (i.e., $\alpha_{ij}^* = 0 \Leftrightarrow \rho_n = \rho_c$), direct-top forecasts result in lower (higher) forecast errors than bottom-up forecasts when ρ_n is negative (positive).*

The arguments to support this hypothesis are based on two assumptions that served to simplify the exposition: a linear relationship between random judgment error and time series uncertainty, and uncorrelated item-level judgment errors. As long as random

[6] The ratio expressed in Equation (9) assumes uncorrelated item-level judgment errors and that these judgment errors scale linearly with the inherent predictability of the time series.

judgment errors are increasing in time series uncertainty, the linearity assumption can be relaxed and the same directional pattern holds: bottom-up forecasting becomes more attractive as one moves from negative to positive ρ_n.[7] Next, there may be positive correlation in the judgment errors if the same forecaster prepares lower-level forecasts. Such positive correlation in judgment errors would dilute the potential benefit of bottom-up forecasting, but the directional pattern expressed in Hypothesis 1 should continue to hold. As we will see in the next section, our experimental design includes a treatment to control for correlation in judgment errors.

3.2. Experimental Design

To test Hypothesis 1, we implement direct-top and bottom-up forecasting tasks in a laboratory setting. Subjects observe 80 periods of demand history prior to making their first forecast(s) and afterward sequentially prepare one-period-ahead forecasts for 30 consecutive periods. Our experimental design in this study follows the main diagonal from Figure 2. We ran one condition with no change or noise correlation across items (0), one condition with positive change and noise correlation (II), and one condition with negative change and noise correlation (IV). Table 1 shows that the top-level series uncertainty ($\ddot{\Sigma}$, defined in Equation (9b)) increases for positive correlations (127.81 in condition II) and decreases for negative correlations (20.47 in condition IV) relative to environments with uncorrelated demand (91.53 in condition 0). In addition to the three demand conditions summarized in Table 1, we administer four different treatments in Study 1. Three of these treatments correspond to different bottom-up procedures, distinguished by the labels BASE, INDEPENDENT, and FEEDBACK; the fourth treatment corresponds to direct-top forecasting (DIRECT-TOP). Besides providing important experimental controls for testing our hypothesis, these four treatments map into different organizational structures.

In the BASE treatment, participants prepare forecasts in both lower-level time series without seeing the implications of their decisions for the aggregate time series, and without being rewarded for their performance in aggregate time series forecasting. This treatment corresponds to a setting where lower-level and top-level forecasting is organizationally separated; however, the same forecaster is responsible for both lower-level forecasts. For example, the same planner may be responsible for forecasting multiple products in the same product family.

[7] For example, if $\sigma_i = (m\Sigma_i)^a$, then $\ddot{\sigma}^{DT}/\ddot{\sigma}^{BU} = 2^{((a-1)/2)}(1+\rho_n)^{a/2}$, and the same directional pattern holds.

Table 1 Experimental Conditions in Study 1

	0	II	IV
Lower-level n and c	40.00	40.00	40.00
Change correlation ρ_c	0.00	0.95	−0.95
Noise correlation ρ_n	0.00	0.95	−0.95
Normative α^*_{ii}	0.62	0.62	0.62
Normative α^*_{ij}	0.00	0.00	0.00
Normative $\ddot{\alpha}^*$	0.62	0.62	0.62
Bottom-up $\Sigma(=\sqrt{2}\times\Sigma_i)$	91.53	91.53	91.53
Top-level noise $(=\sqrt{2}\times n\times\sqrt{(1+\rho_n)})$	56.57	78.99	12.65
Top-level change $(=\sqrt{2}\times c\times\sqrt{(1+\rho_c)})$	56.57	78.99	12.65
Top-level uncertainty, $\ddot{\Sigma}$	91.53	127.81	20.47
Predicted better process	Neither	Bottom-up	Direct-top

In the INDEPENDENT treatment, subjects prepare a forecast for only one lower-level series (while being able to observe the other series). After we collected all data, we generated aggregate forecasts by randomly matching two subjects—one preparing forecasts for the first series and the other preparing forecasts for the second series. This treatment is important as an experimental control, because item-level judgment errors are uncorrelated by construction. Furthermore, this treatment corresponds to a situation in practice where lower-level forecasts across series, as well as lower- and top-level forecasts are organizationally separated. This situation may arise when planners are organized regionally and their (regional) forecasts are combined to create a global forecast.

In the FEEDBACK treatment, subjects prepare forecasts as in the BASE treatment; however, they can see the implied aggregate forecast that results from their lower-level forecasts. This corresponds to an organizational setting where lower and top-level forecasting are organizationally integrated and performed by the same person. Executives engaged in sales and operations planning with whom we corresponded during the design of our study mentioned that this situation often happens—lower-level forecasts sometimes do not "feel right" when aggregated and are revised accordingly. From a methodological perspective, this treatment also has the advantage that we can incentivize participants that make item-level forecasts, according to their performance at the top level.

Our final treatment (DIRECT-TOP) corresponds to direct top-level forecasting. Participants only see the top-level time series and prepare their forecasts for this time series directly. Table 2 contains an overview of these treatments. See Table A.1 in the appendix for a precise breakdown of subjects into treatments, conditions, and data sets across both of our studies.

We use six different data sets in each condition resulting in a 2 (conditions) × 4 (treatments) × 6 (data sets) between-subjects design. Because of subject pool limitations, only two data sets were used in condition IV. Ideally, differences across demand conditions

Table 2 Treatments Used in Study 1

Treatment	Subjects see	Subjects forecast	Top-level forecast	Incentive: Average of	Forecasts/ Subject	No. of subjects
BASE	D_1, D_2	F_1, F_2	$F_1 + F_2$	$MAE(D_1, F_1)$, $MAE(D_2, F_2)$	2×30	173
INDEPENDENT	D_1, D_2	F_1 or F_2	$F_1 + F_2$, From random subjects	$MAE(D_1, F_1)$ or $MAE(D_2, F_2)$	30	144
FEEDBACK	$D_1, D_2, \ddot{D}$	F_1, F_2	$F_1 + F_2$	$MAE(D_1, F_1)$, $MAE(D_2, F_2)$, $MAE(D_1 + D_2, F_1 + F_2)$	2×30	131
DIRECT-TOP	$\ddot{D}$	$\ddot{F}$	$\ddot{F}$	$MAE(\ddot{D}, \ddot{F})$	30	149

are driven entirely by the difference in correlation structure. To create data sets that, between conditions, minimize the amount of sampling noise and emphasize the differences due to the correlation structure, we used the same underlying change and noise random draws to compose the time series presented to subjects. Specifically, for each data set, we generate 2(items) × 2(change, noise) × 110(periods) standard normal draws that are scaled according to the appropriate correlation coefficients for each condition to create the values for $\boldsymbol{\eta}_t$ and $\boldsymbol{\varsigma}_t$ in Equation (1). By appropriately scaling these change and noise random draws, we were able to keep the first time series in each data set (which we refer to as the "blue" series) the same across demand conditions. The second time series for each data set and condition combination (the "red" series) was obtained by scaling the underlying change and noise draws so that the desired variances and correlations were achieved. Keeping the blue time series constant across different correlation structures facilitates comparisons across conditions. Figures 1 and 2 visualize this aspect of our experimental design; the lower of the two lower-level series is the same throughout all conditions.

The forecasting task was implemented in the experimental software z-Tree (Fischbacher 2007); a screenshot of the interface is given in Figure A.1 in the appendix. The study was conducted at the Laboratory for Economic Management and Auctions at the Pennsylvania State University, Smeal College of Business. We recruited subjects from several sections of an undergraduate supply chain core course. To incentivize accurate forecasting, we used a between-subject random incentive system (Schunk and Betsch 2006, Langer and Weber 2008). One student per course section (on average, 100 students) was selected randomly and paid a substantial cash amount based on his or her forecast performance during the experiment. Specifically, each randomly selected student was paid a base payment B minus his or her *mean absolute error* (*MAE*), calculated by averaging all relevant absolute forecast errors across all $T = 30$ periods (see Table 2). The base payment was adjusted depending on treatment and condition, such that the expected earning was roughly equal across treatments and conditions.[8] Across all experimental session reported in this paper (Studies 1 and 2, not including the practitioner and MBA treatments), 48 subjects were randomly selected for payment. The average earning among these subjects was \$81.19, with a standard deviation of \$18.56. Sessions lasted between 30 and 45 minutes.

3.3. Results

Hypothesis 1 is based on the idea that random judgment error is increasing in the uncertainty inherent in the data series. We first examine this behavioral foundation of Hypothesis 1, which requires a measurement of the magnitude of random judgment error in forecasts. We conceptualize random judgment error as a random term appended to an otherwise deterministic forecasting rule and estimate behavioral forecasting models on the lower-level blue series within each condition across the bottom-up treatments BASE, INDEPENDENT, and FEEDBACK. For each subject s, our data comprise forecasts $F_{s,t}$ and demands $D_{d,t}$ for $T = 30$ consecutive periods ($t = 1..T$), with the understanding that demand observations are nested in a particular data set d. Furthermore, let $\Delta F_t = F_t - F_{t-1}$, and let $\Delta D_t = D_t - D_{t-1}$. The forecast errors in a focal series are denoted by $FE_{s,t}$, and the forecast errors in a distal series are denoted by $DE_{s,t}$. The resulting specification is as follows:

$$\begin{aligned} F_{s,t+1} = {} & a_0 + a_1^{d,s} FE_{s,t} + a_2^{d,s} DE_{s,t} + a_3 F_{s,t} + a_4^{d,s} \Delta D_{d,t} \\ & + a_5 \Delta D_{d,t-1} + a_6 \Delta F_{s,t} + a_7 \Delta F_{s,t-1} \\ & + u_d + v_s + \varepsilon_t. \end{aligned} \tag{10}$$

This model includes random effects by data set (u_d) and subject (v_s). It regresses previous forecasts and demands on current forecasts and allows for the possibility that forecasts are influenced by distal series forecast errors in addition to focal errors. The specification is mostly similar to that of Kremer et al. (2011) and accounts for the variance in observed forecasts

[8] For example, for demand condition 0, the base payment was $B =$ \$145 for treatments BASE and INDEPENDENT, $B =$ \$160 for treatment FEEDBACK, and $B =$ \$180 for treatment DIRECT-TOP.

that is due to forecast errors, recent previous forecasts, and recently observed demand. For the DIRECT-TOP treatment, we estimated the same model, but omitted the term $DE_{s,t}$. Estimation results for all treatments and conditions used in this paper are summarized in Table A.2 in the appendix.

We calculated the residual errors for each model. Because the models contain random effects at the subject level, the regression residuals capture within-subject variance but not between-subject variance. For DIRECT-TOP forecasting, we calculated the standard deviation of these residuals as a measure of the magnitude of random judgment error. For all bottom-up treatments, we summed up the resulting lower-level residuals and calculated the standard deviation of this sum as a measure of random judgment error in bottom-up forecasting. We refer to this latter estimate as an implied standard deviation. Table 3 contains an overview of this analysis. Note that the INDEPENDENT treatment plays a special role. In the two other bottom-up treatments, lower-level forecasts are prepared by the same person, possibly creating positively correlated random judgment errors, which, as discussed in §3.1, would imply lower benefits of aggregation. In the INDEPENDENT treatment, the two forecasts are created by different people, naturally limiting the potential for correlated random judgment error.

The results from our analysis support the behavioral foundation of Hypothesis 1. The random judgment errors are similar across treatments in condition 0. In condition II, DIRECT-TOP exhibits more random judgment error than any of the bottom-up treatments. In condition IV, the opposite holds; the random judgment error in DIRECT-TOP forecasting is the lowest among all treatments. Note also that the INDEPENDENT treatment in condition II exhibits no increase in random judgment error compared to condition 0 (44.79 versus 50.19), whereas the random judgment error in the other two bottom-up treatments (BASE and FEEDBACK) is much higher in condition II (65.99 and 62.96) than in condition 0 (53.07 and 53.02, respectively). This would imply that lower-level errors are positively correlated in these two treatments, limiting the benefits of random judgment error aggregation obtained by following a bottom-up process. Furthermore, note that cross-condition comparisons of random judgment error with respect to condition IV are invalid, since condition IV is based on estimations of only two data sets (instead of six data sets in the other conditions).

We now test Hypothesis 1 directly by comparing MAEs across treatments. We estimate regression models using individually aggregated MAEs (i.e., averaged at the top level across 30 time periods per subject) as the dependent variable. Independent variables are factor variables for condition, treatment, and data set, as well as all two-way interactions between these variables. Because of heteroskedasticity in our data, Huber–White standard errors are reported and used for significance tests. The resulting predicted marginal means (correcting for imbalances in data sets across treatments and conditions), as well as significance tests for contrasts across these means, are reported in Table 4.

Results provide support for Hypothesis 1. DIRECT-TOP forecasting outperforms all indirect treatments in condition IV, as predicted. In condition II, the MAEs under DIRECT-TOP forecasting are worse than the MAEs under bottom-up forecasting, particularly in

Table 3 Random Judgment Error Across Treatments

Treatment	Condition	$\ddot{\Sigma}$ (theor.)	Random judgment error
BASE ($\ddot{\sigma}^{BU}$, implied)	0 ($N = 87$)	91.53	53.07 (1.10)
INDEPENDENT ($\ddot{\sigma}^{BU}$, implied)	0 ($N = 38$)	91.53	50.19 (1.57)
FEEDBACK ($\ddot{\sigma}^{BU}$, implied)	0 ($N = 52$)	91.53	53.02 (1.42)
DIRECT-TOP ($\ddot{\sigma}^{DT}$, est.)	0 ($N = 51$)	91.53	49.79 (1.00)
BASE ($\ddot{\sigma}^{BU}$, implied)	II ($N = 65$)	127.81	65.99 (1.58)
INDEPENDENT ($\ddot{\sigma}^{BU}$, implied)	II ($N = 22$)	127.81	44.79 (1.84)
FEEDBACK ($\ddot{\sigma}^{BU}$, implied)	II ($N = 54$)	127.81	62.96 (1.65)
DIRECT-TOP ($\ddot{\sigma}^{DT}$, est.)	II ($N = 61$)	127.81	78.34 (1.68)
BASE ($\ddot{\sigma}^{BU}$, implied)	IV ($N = 21$)	12.65	81.09 (2.41)
INDEPENDENT ($\ddot{\sigma}^{BU}$, implied)	IV ($N = 12$)	12.65	83.08 (3.28)
FEEDBACK ($\ddot{\sigma}^{BU}$, implied)	IV ($N = 25$)	12.65	68.26 (1.86)
DIRECT-TOP ($\ddot{\sigma}^{DT}$, est.)	IV ($N = 37$)	12.65	12.43 (0.28)

Notes. The theoretical $\ddot{\Sigma}$ does not represent a prediction of the level of estimated random errors; only a linear relationship between $\ddot{\Sigma}$ and the estimated values is expected. Standard errors of estimates are noted in parentheses next to an estimate.

Table 4 Mean Absolute Errors in Study 1

Treatment	Condition	MAE (normative)	MAE (observed)
BASE	0 ($N = 87$)	70.94	84.03* (1.36)
INDEPENDENT	0 ($N = 38$)	70.94	82.92** (1.61)
FEEDBACK	0 ($N = 52$)	70.94	84.39* (1.69)
DIRECT-TOP	0 ($N = 51$)	70.94	88.69 (1.53)
BASE	II ($N = 65$)	97.93	120.44 (1.63)
INDEPENDENT	II ($N = 22$)	97.93	111.69** (1.85)
FEEDBACK	II ($N = 54$)	97.93	116.50** (1.61)
DIRECT-TOP	II ($N = 61$)	97.93	124.38 (2.04)
BASE	IV ($N = 21$)	17.59	45.35** (3.86)
INDEPENDENT	IV ($N = 12$)	17.59	59.05** (4.18)
FEEDBACK	IV ($N = 25$)	17.59	39.27** (4.16)
DIRECT-TOP	IV ($N = 37$)	17.59	19.22 (0.60)

Notes. N refers to sample size. Effective sample sizes are smaller in the INDEPENDENT treatment since two observations needed to be combined for an aggregate forecast. A significant effect indicates that the MAE in this particular treatment is significantly different from the MAE in the DIRECT-TOP treatment. Tests are based on t-statistics calculated from contrasting marginal means across treatments. Values in parentheses are standard errors.

*$p \leq 0.05$; **$p \leq 0.01$.

Table 5 MAEs under Decision Support in Study 1

Treatment	Condition	MAE (normative)	MAE (observed, non-DS)	MAE (observed, DS)
BASE	0 ($N = 54/28$)	80.31	95.36 (1.99)	88.35 (1.37)
INDEPENDENT	0 ($N = 18/8$)	80.31	95.01 (2.62)	91.99 (1.99)
FEEDBACK	0 ($N = 17/20$)	80.31	90.19* (2.20)	87.36† (2.12)
DIRECT-TOP	0 ($N = 18/21$)	80.31	98.44 (2.68)	92.62 (2.45)
BASE	II ($N = 28/33$)	120.89	144.21 (2.28)	134.11 (2.25)
INDEPENDENT	II ($N = 12/13$)	120.89	136.77* (2.23)	130.66 (1.77)
FEEDBACK	II ($N = 20/25$)	120.89	138.78 (2.42)	132.86 (2.64)
DIRECT-TOP	II ($N = 20/24$)	120.89	144.07 (3.11)	135.15 (2.35)

Notes. A significant effect indicates that the MAE in this particular treatment is significantly different from the MAE in the DIRECT-TOP treatment. Reported sample sizes (N) are for non-DS/DS samples.
†$p \leq 0.10$; *$p \leq 0.05$.

the INDEPENDENT treatment. Note, though, that the same statement holds true for condition 0, which was not predicted. However, performance advantages of DIRECT-TOP seem to be larger in condition II than in condition 0. In particular, consider that in condition 0, the MAE under the INDEPENDENT treatment is 82.92, compared to an MAE of 88.69 under DIRECT-TOP. This represents a 6% reduction in MAE, or, given that the normative MAE is 70.94, a 32% reduction of the performance loss due to judgmental forecasting compared to optimal forecasting. In condition II, a similar comparison yields a 10% reduction in MAE, and a 48% reduction in performance loss due to judgmental forecasting. In other words, the effect size of using bottom-up as opposed to direct-top procedures is almost twice as high in condition II as it is in condition 0.

3.4. Robustness

We subject our analysis to a robustness test by rerunning two of our conditions (0 and II) in a separate treatment where subjects receive decision support. Many demand planners in practice have access to a quantitative model that suggests a forecast to them. Our robustness test establishes whether such access to (useful) decision support changes forecasting performance and the main results reported above. To simulate the influence of decision support on judgment in our context, we ran an additional treatment that offered subjects a system-generated forecast before they had to enter their own forecast. The system-generated forecast corresponded to the optimal univariate forecast for a series. The forecasts were presented to subjects as statistical forecasts based on historical data without any additional detail as to how they were generated. This feature of our design resembles the fact that system-generated forecasts often appear as a black box to decision makers, since the underlying algorithm is not well explained (or understood by decision makers). Furthermore, since forecasters in practice typically have a better understanding of their forecasting context, we also added to the instructions a short qualitative description of the correlation structure between the series. We will refer to this treatment as the decision support (DS) treatment.

An additional 193 subjects were recruited for this robustness test. Because of the smaller number of subjects in this test compared to our original study, only two data sets (instead of six) from our original study were used, and we did not conduct tests in condition IV. The experimental protocol otherwise remained similar to Study 1. For a fair comparison, these new data are only compared to the original data from Study 1 for subjects receiving the same two data sets. We then estimate a similar regression model as in Study 1, adding a variable for decision support and all two-way interactions between decision support, condition, treatment, and data set. Results from our comparison of predicted marginal MAEs are summarized in Table 5.

In general, the MAEs in the DS treatment are lower than in the non-DS treatment, which was expected because subjects were provided an optimal system-generated forecast. As with the non-DS sample, DIRECT-TOP in the DS sample records a higher MAE than the bottom-up treatments for both conditions 0 and II. Although the MAE differences are not significant, for now we remain cautious about the interpretation that decision support mitigates the relative advantages of direct versus indirect forecasting processes, given that the sample sizes in some of our experimental cells are relatively low (e.g., only 13 observations in the INDEPENDENT treatment of condition II under DS) and in light of further evidence on treatment DS that we will report in the context of Study 2 (§4.4) below.

4. Study 2: Tunnel Vision and the Aggregation of Forecasts

To focus on the implications of random judgment errors (and their aggregation), Study 1 considered

demand conditions requiring univariate item-level forecasting (i.e., $\alpha^*_{ij} = 0$). Study 2 is designed to address biases (and their aggregation) that arise specifically in demand conditions with correlation structures that require multivariate forecasting on the lower level (i.e., $\alpha^*_{ij} \neq 0$).

4.1. Hypothesis Development

A key insight from §2 is that if $\rho_c \neq \rho_n$, the optimal forecasting process at the lower level is multivariate. In other words, the correlation structure among demand time series can require forecasters to consider a distal series when preparing a focal forecast (i.e., $\alpha^*_{ij} \neq 0$). Yet making such holistic assessments is challenging. We propose that decision makers tend to prepare their forecasts in a univariate fashion, rather than incorporating information embedded in related time series. Underlying this proposition is the observation that forecasters face a difficult task when preparing forecasts for multiple time series simultaneously: they must not only *detect* the correlation among the lower-level time series change and noise components, but also understand how to *exploit* such correlation.

There is a rich body of literature on the ability of humans to *detect* correlation between continuous random variables, often in the task context of interpreting scatterplots (Lane et al. 1985, Doherty et al. 2007). A central, and robust, finding in this literature is that decision makers underestimate the absolute value of that correlation. Relative to this research, detecting covariation in our forecasting context is more difficult—subjects face a "moving target," and the mere fact that data are presented in a time series format has behavioral implications relative to a task where data points are not temporally related (as in scatterplots). For example, time series data lend themselves to the perception of illusionary trends (DeBondt 1993). Since detecting correlations is challenging for decision makers even in simple contexts (e.g., scatterplot interpretation), we expect that decision makers in our context have even more difficulty in adequately detecting a correlation between time series change and noise components; if such correlations are detected at all, we expect them to be underestimated.

Even if decision makers can identify a correlation between time series components, it is not clear that they can adequately *exploit* that information. Deciding whether to react to an observed forecast error is akin to accepting or rejecting the hypothesis that the forecast error is due to change (rather than noise) in the series. The benefits of multivariate forecasting stem from the fact that the distal series error provides information about this hypothesis test if change and noise correlations are different ($\rho_c \neq \rho_n$). Decision makers, however, tend to be more selective when testing hypotheses. This phenomenon of selective hypothesis testing (Sanbonmatsu et al. 1998) refers to the idea that when multiple hypotheses should be tested (and relevant data exist for each), decision makers tend to follow a sequential approach to testing these hypotheses instead of pursuing a holistic approach. Complementary hypotheses are thereby treated as independent, and decision makers ignore distal evidence in favor of evidence directly linked to a focal hypothesis. The implication for our task context is that the forecaster predominantly uses the error in the focal series to support (or refute) the hypothesis that a large forecast error represents change in the time series (rather than noise), while ignoring the distal series error. We therefore expect that under conditions where forecasters should react to distal series errors and forecast in a multivariate fashion, they will largely fail to do so, i.e., ($\hat{\alpha}_{ij} = 0 \leq |\alpha^*_{ij}|$). We term this "tunnel vision."

Establishing that the information-processing capabilities of forecasters are insufficient to process multivariate time series information is not the primary purpose of our study. Rather, our objective is to establish the implications of this insight for the comparison of direct-top versus bottom-up forecasting. Recall from Equation (6) the link between optimal direct-top and bottom-up forecasting, i.e., $\ddot{\alpha}^* = \alpha^*_{ii} + \alpha^*_{ij}$, which highlights the important role of the cross-item error-response parameter α_{ij}. First, consider demand conditions for which distal lower-level series contain no value for focal lower-level forecasts (i.e., $\alpha^*_{ij} = 0$). Because this implies that top-level and lower-level series are structurally equivalent ($\ddot{\alpha}^* = \alpha^*_{ii}$), we would not expect any performance differences due to the possible misreaction to forecast errors; although it is possible that forecasters systematically over- or underreact to their forecast errors ($\hat{\alpha}_{ii} \neq \alpha^*_{ii}$, $\hat{\ddot{\alpha}} \neq \ddot{\alpha}^*$), there is no reason to expect a reaction pattern that is different for bottom-up versus direct-top forecasting (i.e., we would expect $\hat{\alpha}_{ii} = \hat{\ddot{\alpha}}$). However, this expectation changes under demand conditions for which optimal lower-level forecasts incorporate information from distal time series (i.e., $\alpha^*_{ij} \neq 0$). If lower-level forecasts neglect distal series information ($\hat{\alpha}_{ij} = 0$), the resulting bottom-up forecasts implicitly apply the focal lower-level smoothing parameter $\hat{\alpha}_{ii}$ to the top-level data. In contrast, if a forecaster predicts the top-level series directly, her smoothing parameter will adjust to the characteristics of the top-level series. The implication is that tunnel vision benefits direct-top forecasts. Recall, however, from Study 1 that we expect direct-top and bottom-up processes may differ in loss due to random judgment error. For conditions where the anticipated random judgment error loss is similar in direct-top and bottom-up processes, we hypothesize the following:

Table 6 Demand Conditions in Study 2

Condition	0	I	III
Lower-level n and c	40.00	40.00	40.00
Change correlation ρ_c	0.00	0.95	−0.95
Noise correlation ρ_n	0.00	−0.95	0.95
Normative α^*_{ii}	0.62	0.56	0.56
Normative α^*_{ij}	0.00	0.41	−0.41
Normative $\ddot{\alpha}^{\times *}$	0.62	0.98	0.15
Top-level noise $(=\sqrt{2}\times n\times\sqrt{(1+\rho_n)})$	56.57	12.65	78.99
Top-level change $(=\sqrt{2}\times c\times\sqrt{(1+\rho_c)})$	56.57	78.99	12.65
Predicted advantage	Neither	Direct-top	Direct-top

Hypothesis 2 (Tunnel Vision). *If optimal lower-level forecasting is multivariate (i.e., $\alpha^*_{ij} \neq 0$), direct-top forecasts result in lower forecast errors than bottom-up forecasts.*

4.2. Experimental Design and Implementation

To test Hypothesis 2, we create two different demand conditions with varying correlation structure at the lower level. Table 6 provides an overview of these conditions (I and III) and also indicates where we expect benefits of direct-top forecasting according to Hypothesis 2. We include condition 0 from Study 1 as a benchmark, for which our theoretical developments predict that neither forecasting process has a performance advantage. Although our empirical observations from Study 1 (Table 4) suggest that bottom-up procedures may outperform direct-top procedures in condition 0, the important prediction here in Table 6 is that relative performance shifts in favor of direct-top procedures as one moves toward conditions I and III.

4.3. Results

We first examine the behavioral foundations of Hypothesis 2. To that purpose, we compare estimated behavioral smoothing parameters ($\hat{\alpha}$) across conditions and treatments. We estimate Equation (10) on the lower-level blue series within each condition across the bottom-up treatments BASE, INDEPENDENT, and FEEDBACK. Across all treatment and condition combinations where the optimal forecasting process requires reacting to distal series error, the distal series' forecast errors had no significant effect on focal series' forecasts. Furthermore, there was no evidence that forecast accuracy within the blue series improved under conditions I or III where distal-item correlation offers theoretical potential for improvement. Together, these observations support our assertion that forecasters suffer from tunnel vision.

To more directly test the foundations of Hypothesis 2, we create estimates of the sum of focal and distal series behavioral smoothing parameters (although the latter were basically equivalent to zero) and record this value as an implied bottom-up smoothing parameter ($\hat{\alpha}_{ii} + \hat{\alpha}_{ij}$). We then estimate behavioral smoothing parameters in the DIRECT-TOP treatment directly and compare these to implied smoothing parameters from other treatments in a demand condition. Results from this analysis are summarized in Table 7. It is apparent that the behavioral α's estimated in the DIRECT-TOP treatment change with the characteristics of the time series. Specifically, there is a statistical difference between the DIRECT-TOP behavioral α in conditions 0 and I (0.64 versus 0.82, $p \leq 0.01$) and in conditions 0 and III (0.64 versus 0.42, $p \leq 0.01$). However, the implied α's from all bottom-up treatments appear "locked in" at a value of around 0.70, due to the inability to adjust lower-level forecasts to errors made in the distal time series. More precisely, the overall average implied α across all bottom-up treatments and conditions in Table 7 is 0.71, and none of the nine implied α's is statistically different from this overall average at $p \leq 0.05$. Although the behavioral α's in the DIRECT-TOP condition still show some misreaction (i.e., underreaction in condition I, overreaction in condition III), they are generally closer to the normative benchmark than their implied lower-level bottom-up counterparts. These observations corroborate the logic underlying Hypothesis 2.

Table 7 Comparisons of Smoothing Parameters

Treatment	Condition	Normative ($\ddot{\alpha}^*$)	Estimated
BASE	0 ($N = 87$)	0.62	0.67 (0.04)
INDEPENDENT	0 ($N = 38$)	0.62	0.72 (0.05)
FEEDBACK	0 ($N = 52$)	0.62	0.73 (0.04)
DIRECT-TOP	0 ($N = 51$)	0.62	0.65 (0.04)
BASE	I ($N = 61$)	0.98	0.73 (0.03)
INDEPENDENT	I ($N = 35$)	0.98	0.66 (0.06)
FEEDBACK	I ($N = 61$)	0.98	0.75 (0.04)
DIRECT-TOP	I ($N = 57$)	0.98	0.80 (0.04)
BASE	III ($N = 63$)	0.15	0.64 (0.05)
INDEPENDENT	III ($N = 34$)	0.15	0.75 (0.05)
FEEDBACK	III ($N = 54$)	0.15	0.71 (0.04)
DIRECT-TOP	III ($N = 58$)	0.15	0.43 (0.05)

Note. Standard errors of estimates are noted in parentheses next to an estimate.

We now test Hypothesis 2 directly by examining forecasting performance in the different conditions and treatments. Our hypothesis suggests that direct-top forecasting can outperform bottom-up forecasting in conditions I and III. We prepared the data by calculating the observed absolute errors for every observation and aggregating these numbers across period for each individual. We then ran regression models within each condition (0, I, III) using treatments and data sets (as well as treatment × data set interactions) as explanatory variables. Mean absolute errors were derived from the regression estimation using predicted marginal means across treatments, accounting for imbalances in our sample. All analyses were run in STATA 13.1. Standard errors were

Table 8 Mean Absolute Error Comparison in Study 2

Treatment	Condition	MAE (normative)	MAE (observed)
BASE	0 ($N = 87$)	70.94	84.03* (1.36)
INDEPENDENT	0 ($N = 38$)	70.94	82.92** (1.61)
FEEDBACK	0 ($N = 52$)	70.94	84.39* (1.69)
DIRECT-TOP	0 ($N = 51$)	70.94	88.69 (1.53)
BASE	I ($N = 61$)	60.17	74.77* (1.50)
INDEPENDENT	I ($N = 35$)	60.17	75.41* (2.15)
FEEDBACK	I ($N = 61$)	60.17	71.98 (1.39)
DIRECT-TOP	I ($N = 57$)	60.17	70.64 (1.34)
BASE	III ($N = 63$)	70.90	92.39** (1.14)
INDEPENDENT	III ($N = 34$)	70.90	95.88** (1.93)
FEEDBACK	III ($N = 54$)	70.90	92.53* (2.14)
DIRECT-TOP	III ($N = 58$)	70.90	86.34 (1.64)

Notes. N refers to sample size. A significant effect indicates that the MAE in this particular treatment is significantly different from the MAE in the DIRECT-TOP treatment. Tests are based on t-statistics calculated from contrasting marginal means across treatments. Values in parentheses are standard errors.

*$p \leq 0.05$; **$p \leq 0.01$.

calculated using Huber–White (i.e., robust) estimators to account for possible heteroskedasticity across treatments and conditions. We further estimated contrasts across treatments comparing bottom-up MAEs to the DIRECT-TOP MAEs. Results from this analysis are detailed in Table 8.

The results in Table 8 are overall consistent with Hypothesis 2. As predicted, in both conditions I and III, DIRECT-TOP forecasting tends to perform better than all bottom-up procedures (BASE, INDEPENDENT, and FEEDBACK). An overall test whether DIRECT-TOP outperforms the pooled sample of all bottom-up treatments reveals a significant performance difference, both in conditions I ($b = 3.15$, $p \leq 0.05$) and III ($b = 6.94$, $p \leq 0.01$). These performance differences occur despite DIRECT-TOP having a performance disadvantage in condition 0 ($b = -4.37$, $p \leq 0.01$). In other words, although we expected DIRECT-TOP to perform similarly to any bottom-up procedure if lower-level series are uncorrelated, DIRECT-TOP appears to be at a disadvantage here. The effect sizes observable in conditions I and III not only offset this apparent natural disadvantage of DIRECT-TOP, but are strong enough to create a performance advantage of DIRECT-TOP instead.

Table 8 also reveals that the three bottom-up treatments perform relatively similarly to each other. We had expected that INDEPENDENT could outperform the BASE treatment, since forecasts across series here are naturally uncorrelated. We had also expected FEEDBACK to outperform the BASE treatment, since observing the implied forecast may create a natural opportunity for error correction. Yet neither of these expected performance advantages is visible in Table 8.

Although the focus of this study was to examine the impact of the tunnel vision *judgment* bias on the choice between a direct or indirect forecasting process, we close by noting that a similar effect may exist in firms relying entirely on *statistical* forecasting. Specifically, most demand planners we interacted with during our study confirmed that their quantitative models are predominantly univariate in nature.[9] If that is the case, statistical forecasts also suffer from "tunnel vision," and direct-top forecasting may provide benefits as well under conditions I and III, even if statistical (as opposed to judgmental) forecasts are involved.

4.4. Robustness

We provide two robustness tests of our experiment. Similar to the first study, we ran a decision support treatment where subjects were provided optimal univariate exponential smoothing forecasts. For the conditions tested in Study 1, these univariate forecasts were optimal. Here, these forecasts are good (optimal in condition 0), but judgmental forecasts could potentially outperform the provided statistical forecasts (in conditions I and III). We collected data from an additional 266 subjects, across conditions (0, I, and III), treatments (BASE, INDEPENDENT, FEEDBACK, and DIRECT-TOP) and two data sets (5 and 6). Table A.1 in the appendix gives a breakdown of the sample used in this study. Similar to our previous analysis, we estimate regression models within each condition and calculate predicted marginal means across treatments. Table 9 provides an overview of our analysis and reports comparable statistics using our previous data set without DS (estimated only on data sets 5 and 6).

Results are consistent between DS and non-DS treatments. In condition III, DIRECT-TOP outperforms all bottom-up treatments ($b = 18.53$, $p \leq 0.01$). Performance under DS appears slightly improved compared to non-DS. In condition I, DIRECT-TOP only barely outperforms the bottom-up treatments ($b = 4.64$, $p \leq 0.10$). In condition 0, DIRECT-TOP again performs slightly worse than the bottom-up treatments ($b = -4.84$, $p \leq 0.05$). Overall, these results are consistent with our previous analysis, indicating that the presence of decision support, although possibly improving performance overall, has little influence on the effects predicted in Hypothesis 2.

Our second robustness test concerns the choice of subject pool. The participants in our previous

[9] Lütkepohl (1984) offers a possible statistical justification for this practice. He shows that, with limited data, univariate autoregressive integrated moving average (ARIMA) models may outperform the theoretically correct multivariate ARIMA model, due to lower parameter estimation error in the univariate models (since fewer parameters need to be estimated).

Table 9 Mean Absolute Error Comparison in Study 2 Robustness Tests

Treatment	Condition (set 5/6)	MAE (normative)	MAE (observed, non-DS)	MAE (observed, DS)
BASE	0 ($N = 54/28$)	80.31	95.36 (1.99)	88.35 (1.37)
INDEPENDENT	0 ($N = 18/8$)	80.31	95.01 (2.62)	91.99 (1.99)
FEEDBACK	0 ($N = 17/20$)	80.31	90.19* (2.20)	87.36† (2.12)
DIRECT-TOP	0 ($N = 18/21$)	80.31	98.44 (2.68)	92.62 (2.45)
BASE	I ($N = 23/30$)	65.94	80.19** (2.18)	80.45* (1.71)
INDEPENDENT	I ($N = 13/8$)	65.94	80.95* (3.60)	84.35** (2.81)
FEEDBACK	I ($N = 16/19$)	65.94	75.19 (2.49)	77.48 (1.95)
DIRECT-TOP	I ($N = 20/24$)	65.94	73.36 (1.71)	74.49 (2.23)
BASE	III ($N = 23/26$)	71.37	105.36** (1.89)	100.71** (2.40)
INDEPENDENT	III ($N = 12/8$)	71.37	104.94** (2.92)	103.44** (2.49)
FEEDBACK	III ($N = 15/15$)	71.37	101.91* (3.69)	99.29** (2.98)
DIRECT-TOP	III ($N = 20/23$)	71.37	87.36 (2.45)	83.59 (2.03)

Note. A significant effect indicates that the MAE in this particular treatment is significantly different from the MAE in the DIRECT-TOP treatment in the decision support sample.
†$p \le 0.10$; *$p \le 0.05$; **$p \le 0.01$.

Table 10 Mean Absolute Error in Study 2 Under Decision Support

Treatment	Condition	MAE (norm.)	MAE (DS)	MAE (MBA)	MAE (PRACT)
BASE	I	52.09	69.36 (3.13)	66.27 (3.11)	66.78* (3.75)
DIRECT-TOP	I	52.09	61.12 (4.10)	60.06 (2.18)	58.73 (2.01)
BASE	III	71.69	103.53** (4.34)	96.23** (1.92)	94.61** (3.02)
DIRECT-TOP	III	71.69	79.28 (1.42)	84.69 (3.51)	83.77 (3.19)

*$p \le 0.05$; **$p \le 0.01$

experiments were undergraduate students. We now report on a targeted, smaller-scale study with subjects having more expertise. Specificaly, we take two treatments from our previous design (BASE and DIRECT-TOP), one data set (5), and two conditions (I and III) and test whether Hypothesis 2 continues to hold in samples comprised of (a) 44 MBA students at a large public university (referred to as the MBA group) and (b) 23 forecasting practitioners from a Fortune 500 U.S. company (referred to as the PRACT group). In this robustness test, subjects received the DS treatment. The more experienced decision makers are therefore only compared to the sample of undergraduate students that received the same treatment ($N = 54$). Furthermore, to gain more information, the forecasting experts in the PRACT group completed the exercise twice—once in the BASE treatment and once in the DIRECT-TOP treatment. We varied the order in which these treatments were administered at random. We then estimated regression models to predict MAEs in each condition and treatment. Among the PRACT sample, we estimated regression models with subject random effects and a fixed effect for the order of treatments. Results are summarized in Table 10. The MAEs are consistent across groups. In condition III, DIRECT-TOP outperforms BASE across all samples. In condition I, although effect sizes are not significant ($p = 0.11$) in the MBA sample (as well as in the original DS sample), DIRECT-TOP also outperforms BASE in the PRACT sample ($p \le 0.05$).

5. Discussion

The goal of our research is to provide guidance as to how judgmental hierarchical forecasting processes should be structured. In this section, we summarize insights from our two studies, illustrate how a firm might apply our findings, and discuss the potential economic impact of our research.

5.1. Summary

Our two studies have highlighted that whether bottom-up or direct-top forecasting is advantageous from a judgmental forecasting perspective depends to a large degree on the underlying correlation structure at the lower level. The aggregation of random judgment errors implies that bottom-up forecasting can become advantageous when data aggregation results in top-level series with much higher inherent uncertainty than the item-level series (for example, when change and noise correlations are positively correlated). Tunnel vision implies that direct-top forecasting becomes more advantageous the more information can be captured by multivariate lower-level forecasting, because bottom-up forecasts largely fail to exploit this information. We have summarized the resulting prescriptive framework in Figure 3.

Figure 3 Prescriptive Framework

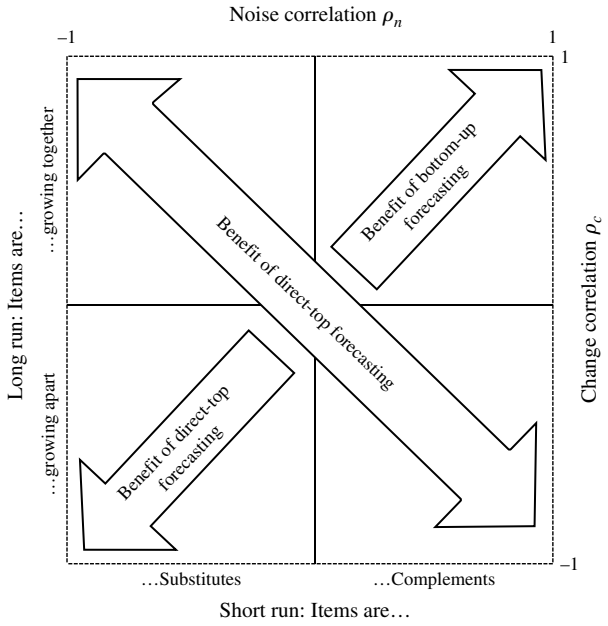
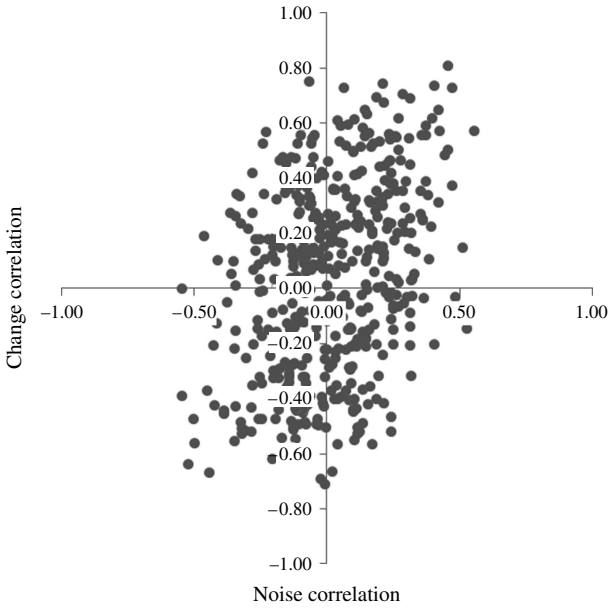


Figure 4 Example Empirical Measurement



The primary implication of Figure 3 is that the process used for forecasting top-level demand should depend on the degree of substitutability/complementarity that exists between item demands. On a very basic level, Figure 3 implies that any concerns of substitutability between products, whether in long-term growth (i.e., change) or short-term effects (i.e., noise), create the potential for direct-top forecasting to be beneficial. In the absence of substitutability concerns, particularly if products are complementary or affected by the same growth drivers, bottom-up forecasting may become beneficial.

5.2. Applications

If a forecaster understands whether substitutability or complementarity is present among products, the choice of which forecasting process to follow is directionally clear from our framework. We were also interested in developing and demonstrating an empirical approach to measuring change and noise correlations between time series. For this, we obtained three years of monthly data in 34 demand time series from a Fortune 500 company, together with the consensus forecasts that were implemented. These consensus forecasts were based on a quantitative forecast, but contained substantial judgmental adjustments. The initial quantitative forecasts were not available for analysis. We therefore used the first two years of the data to fit several generalized exponential smoothing models to each time series (Hyndman et al. 2008) and used the best-fitting model to create forecasts in the remaining holdout year. All analyses were run in R. The statistical models beat the consensus forecast in the holdout sample in 29 of 34 series with a combined MAE that was 23% lower than the MAE resulting from the consensus forecasts, supporting the notion that forecasts in practice are indeed subject to judgmental error.

To illustrate how our insights could be applied, we subjected each of the 34 (real) time series to the following procedure: (1) Time series were deseasonalized using monthly indexes and detrended using estimates of month-to-month growth over three years. (2) For each time series, we estimated a simple state space model in STATA 13.1 to estimate the level of each time series in each period. (3) We used the difference between consecutive level estimates in a time series as an estimate for change in the series, and the difference between demand and level estimate as an estimate for noise in the series. Then, for each pair among these 34 time series, we calculated the correlation coefficient between their change and noise components and mapped these sets of correlation coefficients in Figure 4.

We can see many dyads close to the center of Figure 4 (similar to our condition 0), as well as in the top right quadrant (similar to our condition II), where our empirical results suggest that a bottom-up process may have an advantage. There are also many dyads in those quadrants where our results suggest a performance advantage to a direct-top process (similar to our conditions I, III, and IV).

5.3. Performance Improvements

The MAE estimates in Tables 4 and 8 suggest that by selecting the right forecasting process, MAEs can drop by 5–40 units (comparing INDEPENDENT and TOP-DIRECT within each condition). Averaged across conditions, this represents a 20% reduction in MAE.

Excluding the more extreme condition IV, the average reduction in MAE is still 8%. In terms of *mean absolute percentage error* (*MAPE*), the performance improvements we document are not high—but our study was not designed for interpretation in terms of MAPE. The arbitrary starting values of the item-level time series in our experiments were $\boldsymbol{\mu}_0 = (\mu_{1,0}, \mu_{2,0}) = (2{,}000, 2{,}000)$, which for the top-level time series implied 4,000 units—the MAEs of optimal forecasting were (depending on condition) at around 100, resulting in an optimal MAPE of just 2.5%. Not surprisingly, then, differences in our MAEs are very small when expressed as MAPEs; if the MAE drops by 10 units from 100 to 90, the MAPE drops from 2.5% to 2.25%, i.e., a 10% reduction. Nevertheless, this performance improvement is similar to the performance improvements recorded by Kesavan et al. (2010), who show that moving from consensus forecasting to their statistical model decreases MAPE on average from 4.40% to 4.09%. These numbers are also close to the statistics presented by Osadchiy et al. (2013), who show an 11% reduction in MAPE of market-based forecasts compared to analysts' forecasts.

5.4. Economic Significance

Numerous studies have explored and established the general link between forecast accuracy and operational performance, both theoretically and empirically. Early simulation research focused on manufacturing environments (e.g., Ritzman and King 1993). More recently, simulations and analytical models have examined the value of information sharing in supply chains by documenting how improved forecast performance due to information sharing translates into improved performance (e.g., Aviv 2001, Zhao et al. 2002). A seminal study using the case of a military distribution system together with inventory system simulation demonstrates how using better forecasting models allows achieving similar service levels with about 7% less investment in inventory (Gardner 1990). In a similar case study/simulation in the context of labor demand forecasting in a warehouse, Sanders and Graman (2009) show that improved forecasting performance can lead to cost performance improvements of up to 40%.

Several empirical case studies demonstrate the effect of changes in particular forecasting process/methods within an organization. Documenting the case of a major overhaul of forecasting at Coca-Cola enterprises, a reduction of the forecast error by, on average, 15% is associated with an approximately 25% reduction in days of inventory (Clark 2006). A similar study of spare parts forecasting at Hewlett Packard shows that a 10% reduction in forecast error leads to an increase in on-time delivery from 60% to 95% (Shan et al. 2009). Furthermore, a study of the overhaul of the forecasting process at an electronics manufacturer documents that a decrease in the forecast percentage errors from 42% to 12% leads to a doubling of inventory turns and a decrease by about 50% in on-hand inventory (Oliva and Watson 2009). In a study of judgmental forecast adjustments, a one-Percentage-point improvement in average absolute percentage forecast error leads to inventory reductions of 15%–20% and increases in the fill rates of about 1% (Syntetos et al. 2010). Similarly, a summary of a supply chain benchmarking study indicates that every one-percentage-point improvement in forecast accuracy leads to a two-percentage-point improvement in fill rates (Hofman 2004). All of these studies show that the relatively low-cost intervention of improving a forecasting process can lead to very high supply chain benefits.

Finally, we apply our MAE savings to the base-stock inventory model developed by Graves (1999) for the same nonstationary demand process we study here. Graves (1999) establishes that the total amount of safety stock required depends on the forecast accuracy, the change-to-noise ratio ("weight") of the time series, the desired service level, and the replenishment lead time. All else equal, a given percentage improvement in forecast error translates to the same percentage reduction in safety stock. So, our 20% MAE reduction would lead to a 20% reduction in safety stock. A key insight from Graves (1999) is that the necessary safety stock for a given service level and lead time is increasing in the weight of the time series. This suggests that forecast improvements are particularly valuable for aggregate series with high weight (and thus high $\ddot{\alpha}^*$). For example, with a lead time of three periods and a 95% target service level, a 10 unit reduction in MAE would lead to a reduction in safety stock of 41 units for our condition I ($\ddot{\alpha}^* = 0.15$) and a 76 unit reduction in safety stock for our condition III ($\ddot{\alpha}^* = 0.98$).

6. Conclusion

Our study is, to the best of our knowledge, the first to examine hierarchical forecasting from a behavioral perspective. Our data show that the accuracy of *judgmental* top-level forecasts is generally not the same for top-*direct* and bottom-up *indirect* processes. We show how the correlation structure of lower-level demands systematically drives relative process performance, through two behavioral phenomena with known fundamentals, but with hitherto unknown implications for hierarchical judgmental forecasting: (a) random judgment errors benefit bottom-up (top-direct) in demand conditions where item-level correlations imply high (low) inherent uncertainty in the top series, and (b) the inability of forecasters to detect and exploit correlation structures in data benefits top-direct forecasts in demand conditions where

item-level correlations imply high value of multivariate forecasting. Although together these mechanisms do not establish a general preference for one process over the other, our research provides a clear framework that maps process preference on elementary properties of the forecasting environment. Bottom-up forecasts have performance advantages if lower-level items are affected similarly by short- and long-term shocks, e.g., products that are affected similarly by general market growth ("change") and weather effects ("noise"). Although this finding resonates well with the established wisdom that bottom-up forecasting processes are generally preferable (e.g., Dangerfield and Morris 1992), our research points to a broad range of demand environments that puts bottom-up procedures at a disadvantage. Top-direct procedures have advantages for lower-level items that are substitutes either in the long run or in the short run.

The resulting framework we develop in Figure 3 is easy to communicate, and thus it provides a good basis for structuring forecasting processes in practice. The applicability of our findings in practice does not hinge on precise knowledge of the structure and parameters of the data-generating process (such as correlation coefficients), but the firm must have a qualitative understanding of the direction of long- and short-term correlations among their demand items. Our experience is that demand planners in practice are generally aware of whether the products for which they plan are substitutes or complements in the short and long run, and as such, will find our framework applicable to their work.

The framework in our study has some limitations. Related to our choices of experimental design and implementation, we assumed (a) equal lower-level change and noise ($c_i = n_i$), (b) symmetry among lower-level time series ($c_i = c_j$, $n_i = n_j$), (c) two lower-level items ($M = 2$), (d) fairly large distal correlations ($|\rho_n| = |\rho_c| = 0.95$), (e) the absence of trends and seasonality, and (f) that forecasts are needed for one period ahead. Each one of these assumptions served the purpose of increasing experimental control and simplifying subjects' task in an already complex environment, and could be relaxed in future studies. Interestingly, one could argue that the performance effects we demonstrate are on the conservative side *because* of our simplifying assumptions. In particular, remember that the structure and parameters of the demand process drive the *value of multivariate forecasting*, which in turn drives the performance loss associated with forecasters inability to capture this value in the context of a bottom-up forecasting process (Hypothesis 2). With regard to (c) and (d) above, it can be shown that the theoretical value of multivariate forecasting tends to increase in the number of lower-level items, even for more moderate between-item correlations than those used in our experimental implementation.[10] With regards to points (e) and (f) above, it stands to reason that either needing to additionally estimate trend and seasonal components or forecasting many periods ahead (as opposed to the one-period-ahead forecasts we required) would increase forecast errors across the board, but would also widen the performance gap between different forecasting processes. Although clearly speculative, these ideas point to important research questions that our study did not address directly.

In general, the time series we use are artificial, and our experiment emphasizes internal validity. Field tests of our framework, using real-world time series and the judgment of experienced forecasters, are needed to strengthen the external validity of this framework. Finally, our analysis focused on the comparison between bottom-up and direct-top procedures for forecasting top-level demand. An equally important issue in the context of sales and operations planning in practice is a comparison of top-down versus direct lower-level procedures for forecasting item-level demands. We have completed such a comparison, but do not report the details from our analysis here. Results are available from the authors upon request. In a nutshell, the performance advantages of direct-top forecasting "trickle down" to the lower level through top-down forecasting if an adequate disaggregation mechanism is used.

Acknowledgments

The authors greatly appreciate the valuable feedback on earlier versions of this manuscript provided by J. Scott Armstrong, Karen Donohue, Robert Fildes, Paul Goodwin, Yun Shin Lee, and Ulrich Thonemann. They are also grateful for the comments made during seminar presentations at Ohio State University, Penn State University, European Business School, University of California, Los Angeles, SDA Bocconi, University College London, the University of Texas at Dallas, Cambridge University, Nanyang Business School, Instituto de Estudios Superiores de la Empresa Barcelona, and École Supérieure des Sciences Économiques et Commerciales Paris, as well as at conference presentations at the Institute for Operations Research and the Management Sciences, Production and Operations Management Society, International Symposium on Forecasting, and the Utah Winter Operations Conference. The authors also acknowledge the constructive feedback provided by the associate editor and three anonymous referees. This research was funded in part by the Smeal College of Business.

[10] For example, with values of $\rho_c = -0.2$ and $\rho_n = 0.2$, and $M = 6$, direct-top forecasting would result in a mean absolute error 20% lower than the mean absolute error obtained by univariate, bottom-up forecasting. Essentially, when a small amount can be learned from each of several other correlated time series, the value of multivariate forecasting can be large. Behaviorally, of course, the task of trying to exploit multiple weaker correlations is significantly more challenging than the $M = 2$ setting explored in our study.

Appendix

Figure A.1 (Color online) Screenshot for Treatment BASE

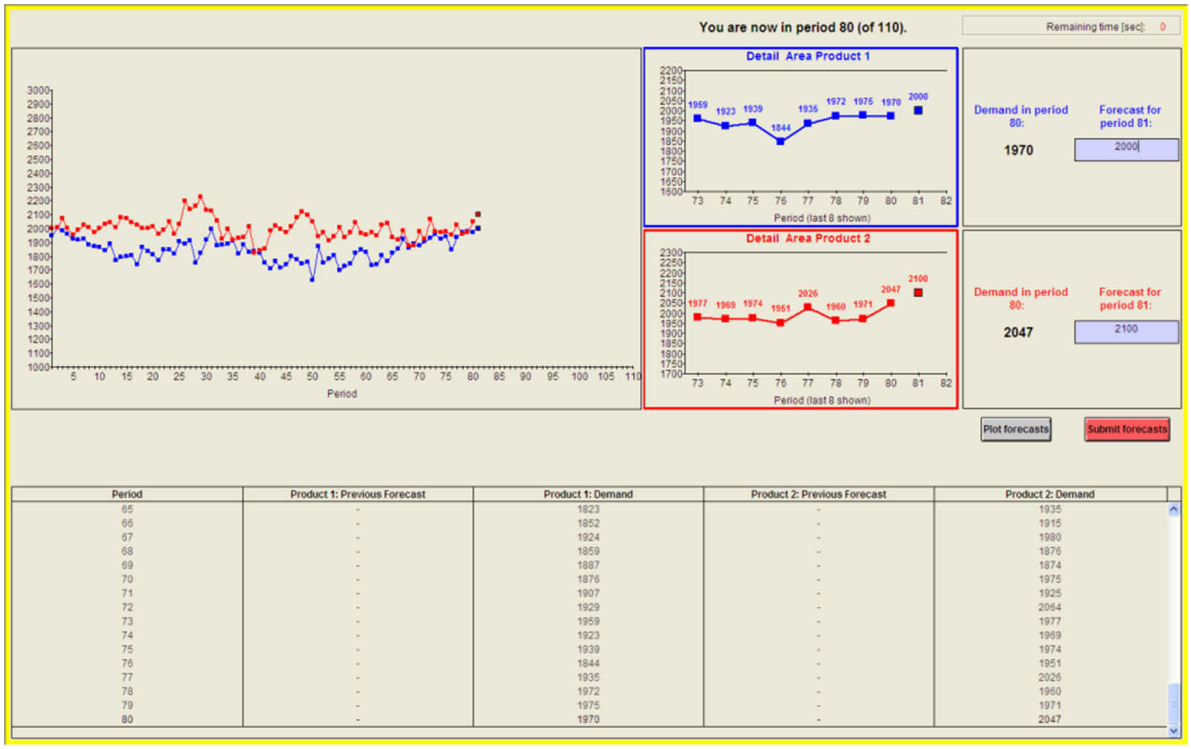


Table A.1 Breakdown of Sample by Treatment, Condition, and Data Set

	BOTTOM-UP																							
Treatment	BASE						INDEPENDENT						FEEDBACK						DIRECT-TOP					
Data set	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Condition 0	8	9	8	8	25	29	5	5	5	5	8	10	9	9	9	8	8	9	8	7	9	9	10	8
Condition I	9	10	10	9	12	11	5	6	5	6	8	5	12	11	11	11	8	8	10	10	10	7	10	10
Condition II	9	10	8	10	15	13	2	2	1	5	6	6	10	10	6	8	10	10	11	10	10	10	10	10
Condition III	11	8	11	10	12	11	5	5	6	6	6	6	9	10	10	10	8	7	10	10	9	9	10	10
Condition IV	0	0	0	0	12	9	0	0	0	0	6	6	0	0	0	0	12	13	0	0	0	0	15	22

Note. Note that in the INDEPENDENT treatment, two subjects were necessary for an observation, which leads to a smaller effective sample size within that treatment.

Table A.2 Behavioral Estimates

Condition	0		I		II		III		IV	
	BASE treatment									
$FE_{s,t}$	0.66**	(0.03)	0.73**	(0.04)	0.77**	(0.05)	0.64**	(0.05)	0.80**	(0.08)
$DE_{s,t}$	0.01	(0.01)	0.00	(0.01)	−0.07*	(0.03)	0.00	(0.01)	0.01	(0.05)
$F_{s,t}$	0.99**	(0.01)	1.00**	(0.00)	0.99**	(0.00)	1.00**	(0.00)	0.96**	(0.03)
$\Delta D_{d,t}$	0.16**	(0.03)	0.05	(0.03)	0.08**	(0.03)	0.15**	(0.03)	−0.05	(0.09)
$\Delta D_{d,t-1}$	0.10**	(0.03)	0.07**	(0.03)	0.09**	(0.02)	0.10**	(0.03)	0.02	(0.07)
$\Delta F_{s,t}$	−0.19**	(0.02)	−0.18**	(0.03)	−0.16**	(0.02)	−0.20**	(0.02)	−0.17**	(0.06)
$\Delta F_{s,t-1}$	−0.09**	(0.01)	−0.06**	(0.01)	−0.09**	(0.01)	−0.08**	(0.01)	−0.11**	(0.04)
Constant	18.62*	(9.09)	10.15	(8.41)	21.06**	(7.66)	8.07	(8.31)	66.44	(58.00)
N	2,349		1,647		1,755		1,701		567	
Subjects	87		61		65		63		21	

Table A.2 (Continued)

Condition	0	I	II	III	IV
			INDEPENDENT treatment		
$FE_{s,t}$	0.72** (0.05)	0.66** (0.04)	0.70** (0.06)	0.75** (0.05)	0.70** (0.07)
$F_{s,t}$	0.99** (0.01)	0.99** (0.01)	1.00** (0.01)	0.99** (0.01)	0.99** (0.03)
$\Delta D_{d,t}$	0.08* (0.04)	0.10** (0.04)	0.12* (0.05)	0.03 (0.04)	0.08 (0.07)
$\Delta D_{d,t-1}$	−0.01 (0.03)	0.08** (0.03)	0.11** (0.04)	0.08* (0.03)	0.02 (0.07)
$\Delta F_{s,t}$	−0.07* (0.03)	−0.22** (0.03)	−0.21** (0.04)	−0.21** (0.03)	−0.17** (0.06)
$\Delta F_{s,t-1}$	−0.09** (0.02)	−0.10** (0.02)	−0.06** (0.02)	−0.08** (0.02)	−0.14** (0.03)
Constant	24.43* (15.47)	24.11* (13.07)	−0.54 (13.35)	15.83 (12.83)	20.12 (46.87)
N	1,026	945	594	918	324
Subjects	38	35	22	34	12
			FEEDBACK treatment		
$FE_{s,t}$	0.72** (0.04)	0.73** (0.04)	0.72** (0.05)	0.70** (0.04)	0.75** (0.08)
$DE_{s,t}$	0.01 (0.01)	0.02 (0.01)	−0.06* (0.03)	0.01 (0.01)	−0.04 (0.02)
$F_{s,t}$	0.99** (0.01)	0.99** (0.00)	0.99** (0.00)	0.99** (0.01)	0.99** (0.02)
$\Delta D_{d,t}$	0.18** (0.04)	0.05 (0.03)	0.13** (0.03)	0.05 (0.03)	0.03 (0.05)
$\Delta D_{d,t-1}$	0.09* (0.04)	0.03 (0.02)	0.10** (0.04)	0.07* (0.03)	0.12 (0.04)
$\Delta F_{s,t}$	−0.18** (0.03)	−0.16** (0.02)	−0.23** (0.02)	−0.14** (0.02)	−0.12** (0.04)
$\Delta F_{s,t-1}$	−0.07** (0.01)	−0.07** (0.01)	−0.12** (0.01)	−0.07** (0.01)	−0.06** (0.02)
Constant	19.85* (10.18)	15.14* (7.25)	17.03* (7.59)	28.91* (11.89)	22.20 (33.71)
N	1,404	1,647	1,458	1,458	675
Subjects	52	61	54	54	25
			DIRECT-TOP treatment		
$FE_{s,t}$	0.65** (0.04)	0.80** (0.05)	0.68** (0.03)	0.43** (0.05)	0.72** (0.09)
$F_{s,t}$	0.99** (0.01)	0.99** (0.00)	0.99** (0.01)	0.66** (0.04)	0.88** (0.02)
$\Delta D_{d,t}$	0.15** (0.03)	0.20** (0.04)	0.07* (0.03)	−0.08** (0.03)	0.03 (0.04)
$\Delta D_{d,t-1}$	0.01 (0.03)	0.03 (0.03)	0.08** (0.03)	0.04 (0.02)	−0.01 (0.04)
$\Delta F_{s,t}$	−0.12** (0.03)	−0.08** (0.03)	−0.15** (0.02)	−0.15** (0.03)	−0.07* (0.03)
$\Delta F_{s,t-1}$	−0.02** (0.01)	−0.07** (0.01)	−0.02** (0.01)	−0.09** (0.02)	−0.03 (0.02)
Constant	50.70* (19.68)	29.88 (17.30)	51.25** (18.72)	1,365.6** (157.19)	461.20 (80.89)
N	1,380	1,539	1,659	1,566	999
Subjects	51	57	61	58	37

Notes. Estimates in the BASE, INDEPENDENT, and FEEDBACK treatments were made on forecasts from the blue series. Estimates in the DIRECT-TOP treatment were made on forecasts from the green series. The term $DE_{s,t}$ from Equation (10) was removed from the estimation for both DIRECT-TOP and INDEPENDENT treatments, since no such data were available in these treatments.

*$p \leq 0.05$; **$p \leq 0.01$.

References

Allen PG, Fildes R (2001) Econometric forecasting. Armstrong JS, ed. *Principles of Forecasting* (Kluwer, Norwell, MA), 303–362.

Allon G, Huang T, Bassamboo A (2013) Bounded rationality in service systems. *Manufacturing Service Oper. Management* 15(2): 263–279.

Andreassen P, Kraus SJ (1990) Judgemental extrapolation and the salience of change. *J. Forecasting* 9(4):347–372.

Armstrong JS (1985) *Long-Range Forecasting* (Wiley, New York).

Armstrong JS (2001) Combining forecasts. Armstrong JS, ed. *Principles of Forecasting* (Kluwer, Norwell, MA), 417–437.

Aviv Y (2001) The effect of collaborative forecasting on supply chain performance. *Management Sci.* 47(10):1326–1343.

Blattberg RC, Hoch SJ (1990) Database models and managerial intuition: 50% model +50% manager. *Management Sci.* 36(8): 887–899.

Bowman EH (1963) Consistency and optimality in managerial decision making. *Management Sci.* 9(2):310–321.

Clark S (2006) Managing the introduction of a structured forecast process: Transformation lessons from Coca-Cola Enterprises Inc. *Foresight* 4:21–25.

Dangerfield BJ, Morris JS (1992) Top-down or bottom-up: Aggregate vs. disaggregate extrapolations. *Internat. J. Forecasting* 8(2):233–241.

DeBondt WFM (1993) Betting on trends: Intuitive forecasts of financial risk and return. *Internat. J. Forecasting* 9(3):355–371.

Doherty ME, Anderson RB, Angott AM, Klopfer DS (2007) The perception of scatterplots. *Perception Psychophysics* 69(7):1261–1272.

Enns PG, Machak JA, Spivey WA, Wrobleski WJ (1982) Forecasting applications of an adaptive multiple exponential smoothing model. *Management Sci.* 28(9):1035–1044.

Fildes R, Petropoulos F (2015) Improving forecast quality in practice. *Foresight* 36:5–12.

Fildes R, Goodwin P, Lawrence M, Nikolopoulos K (2009) Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *Internat. J. Forecasting* 25(1):3–23.

Fischbacher U (2007) z-Tree: Zurich toolbox for ready-made economic experiments. *Experiment. Econom.* 10(2):171–178.

Gardner ES (1990) Evaluating forecast performance in an inventory control system. *Management Sci.* 36(4):490–499.

Gaur V, Kesavan S, Raman A, Fisher ML (2007) Estimating demand uncertainty using judgmental forecasts. *Manufacturing Service Oper. Management* 9(4):480–491.

Graves SC (1999) A single-item inventory model for a nonstationary demand process. *Manufacturing Service Oper. Management* 1(1):50–61.

Harrison PJ (1967) Exponential smoothing and short-term sales forecasting. *Management Sci.* 13(11):821–842.

Harvey AC (1986) Analysis and generalization of a multivariate exponential smoothing model. *Management Sci.* 32(3):374–380.

Harvey N (1995) Why are judgments less consistent in less predictable task situations? *Organ. Behav. Human Decision Processes* 63(30):247–263.

Harvey N, Ewert T, West R (1997) Effects of data noise on statistical judgement. *Thinking and Reasoning* 3(2):111–132.

Hofman D (2004) The hierarchy of supply chain metrics. *Supply Chain Management Rev.* 8(6):28–37.

Hyndman RJ, Ahmed RA, Athanasoloulos G, Shang HL (2011) Optimal combination forecasts for hierarchical time series. *Comput. Statist. Data Anal.* 55(9):2579–2589.

Hyndman RJ, Koehler AB, Ord JK, Snyder RD (2008) *Forecasting with Exponential Smoothing: The State Spate Approach* (Springer, Berlin).

Jones RH (1966) Exponential smoothing for multivariate time-series. *J. Royal Statist. Soc.* 28(1):241–251.

Kesavan S, Gaur V, Raman A (2010) Do inventory and gross margin data improve sales forecasts for U.S. public retailers? *Management Sci.* 56(9):1519–1533.

Kremer M, Moritz B, Siemsen E (2011) Demand forecasting behavior: System neglect and change detection. *Management Sci.* 57(10):1827–1843.

Lane DM, Anderson CA, Kellam KL (1985) Judging the relatedness of variables: The psychophysics of covariation detection. *J. Experiment. Psychol.* 11(5):640–649.

Langer T, Weber M (2008) Does commitment or feedback influence myopic loss aversion? An experimental analysis. *J. Econom. Behav. Organ.* 67(3–4):810–819.

Larrick RP, Soll JB (2006) Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Sci.* 52(1):111–127.

Lawrence M, Goodwin P, O'Connor M, Onkal D (2006) Judgmental forecasting: A review of progress over the last 25 years. *Internat. J. Forecasting* 22(3):493–518.

Lawrence M, O'Connor M (1992) Exploring judgmental forecasting. *Internat. J. Forecasting* 8(1):15–26.

Lee YS, Siemsen E (2015) Task decomposition and newsvendor decision making. Working paper, University of Minnesota, Minneapolis.

Lütkepohl H (1984) Forecasting contemporaneously aggregated vector ARMA processes. *J. Bus. Econom. Statist.* 2(3): 201–214.

Lütkepohl H (2007) *New Introduction to Multiple Time Series Analysis* (Springer, Berlin).

Oliva R, Watson N (2009) Managing functional biases in organizational forecasts: A case study of consensus forecasting in supply chain planning. *Production Oper. Management* 18(2): 138–151.

Osadchiy N, Gaur V, Seshadri S (2013) Sales forecasting with financial indicators and experts' input. *Production Oper. Management* 22(5):1056–1076.

Özer Ö, Zheng Y, Chen KY (2011) Trust in forecast information sharing. *Management Sci.* 57(6):1111–1137.

Ritzman LP, King BE (1993) The relative significance of forecast errors in multistage manufacturing. *J. Oper. Management* 11(1): 51–65.

Sanders N, Graman G (2009) Quantifying costs of forecast errors: A case study of the warehouse environment. *Omega* 37(1): 116–125.

Sanbonmatsu DM, Posovac SS, Kardes FR, Mantel SP (1998) Selective hypothesis testing. *Psychonomic Bull. Rev.* 5(2):197–220.

Schunk D, Betsch C (2006) Explaining heterogeneity in utility functions by individual differences in decision modes. *J. Econom. Psychol.* 27(3):386–401.

Schweitzer ME, Cachon G (2000) Decision bias in the newsvendor problem with a known demand distribution: Experimental evidence. *Management Sci.* 46(3):404–420.

Shan JZ, Ward J, Jain S, Beltran J, Amirjalayer F, Kim YW (2009) Spare-parts forecasting: A case study at Hewlett-Packard. *Foresight* 14:40–47.

Su X (2008) Bounded rationality in newsvendor models. *Manufacturing Service Oper. Management* 10(4):566–589.

Syntetos A, Nikolopoulos K, Boylan JE (2010) Judging the judges through accuracy-implication metrics: The case of inventory forecasting. *Internat. J. Forecasting* 26(1):134–143.

Zhao X, Xie J, Leung J (2002) The impact of forecasting model selection on the value of information sharing in a supply chain. *Eur. J. Oper. Res.* 142(2):321–344.