View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Bradford Scholars

UNIVERSITY of BRADFORD

Library

The University of Bradford Institutional Repository

http://bradscholars.brad.ac.uk

This work is made available online in accordance with publisher policies. Please refer to the repository record for this item and our Policy Document available from the repository home page for further information.

To see the final version of this work please visit the publisher's website. Access to the published online version may require a subscription.

Link to publisher's version: *https://doi.org/10.1016/j.ijforecast.2015.12.010*

Citation: Alvarado-Valencia J, Barrero LH, Onkal D et al (2017) Expertise, credibility of system forecasts and integration methods in judgmental demand forecasting. International Journal of Forecasting. 33(1): 298-313.

Copyright statement: © 2016 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved. Reproduced in accordance with the publisher's self-archiving policy. This manuscript version is made available under the *CC-BY-NC-ND 4.0 license*.

Expertise, credibility of system forecasts and integration methods in judgmental demand forecasting

Jorge A. Alvarado-Valencia[a,*], Lope H. Barrero[a], Dilek Önkal[b], Jack Dennerlein[cd]

[*]Corresponding author. E-mail address: jorge.alvarado@javeriana.edu.co. Tel: +571 3208320 ext. 5301

[a]Department of Industrial Engineering, Pontificia Universidad Javeriana, Cra 7#40-62 Ed. Jose Gabriel Maldonado P.3, Bogotá, Colombia.

[b]Faculty of Business Administration, Bilkent University, 06800 Ankara, Turkey. E-mail address:onkal@bilkent.edu.tr

[c]Department of Physical Therapy, Bouvé College of Health Sciences, Northeastern University, 6 Robinson Hall, 360 Huntington Ave., Boston, MA 02115, United States .E-mail address:j.dennerlein@neu.edu.

[d]Harvard School of Public Health, Boston, MA, United States.

Abstract

Expert knowledge elicitation lies at the core of judgmental forecasting – a domain that fully relies on the power of such knowledge and its integration into forecasting. Using experts in a demand forecasting framework, this work aims to compare accuracy improvement and forecasting performance for three judgmental integration methods. To do so, a field study was conducted with 31 experts in four companies. The methods compared were the *Judgmental Adjustment*, the *50-50 Combination, and* the *Divide-and-Conquer*. Forecaster expertise, credibility of system forecasts and the need to rectify system forecasts were also assessed and mechanisms to perform this assessment were considered. When a) a forecaster's relative expertise was high, b) relative credibility of system forecasts was low, and c) system forecasts had a strong need for correction, *Judgmental Adjustment* improved

accuracy over the other integration methods as well as over the system forecasts. Experts with higher expertise showed a higher adjustment frequency. Our results suggest that *Judgmental Adjustment* promises to be valuable in the long term if adequate conditions of forecaster expertise and credibility of system forecasts are met.

Keywords

Judgmental forecasting, expert selection, expert elicitation methods, credibility of system forecasts

1. Introduction

Forecasts present critical inputs into decision-making processes with experts playing vital roles in bringing specialized knowledge not captured by statistical models. The issue of effectively integrating computer capabilities to model historical patterns with human expertise to monitor and assess contextual information has been attracting vast attention, primarily within the judgmental forecasting domain (Lawrence, Goodwin, Oconnor, & Onkal, 2006). Volatile business dynamics and barriers to accessing reliable domain information make it extremely difficult to rely solely on statistical forecasting methods, particularly in situations such as product demand forecasting, when decision impact is large and uncertainty is high (Sanders & Manrodt, 2003). As a result, expert knowledge needs to be systematically incorporated into the process of demand forecast improvement – a process where expertise plays a key part in today's competitive business settings.

Expert knowledge elicitation poses a number of challenging questions to researchers and practitioners in judgmental demand forecasting as well as across a host of other decision-making domains. These include questions on how the elicited responses may be affected by (i) the choice between different elicitation methods within a specific context (Bolger & Rowe, 2014,2015; Cooke, 1991); (ii) the selection and number of experts (Aspinall, 2010); (iii) experts' personal attributes (Budnitz et al., 1997; Morgan, 2014); as well as (iv) the presentation of relevant information to overcome biases (Martin et al., 2012; Morgan, 2014). Judgmental forecasting context offers a good platform to study such issues with apparently conflicting research findings on the contribution of expertise (Lawrence et al., 2006).

In particular, a comparison of various techniques (i.e., judgmental integration methods) for integrating systems advice and human judgment is an important step in assessing how to improve demand forecasting processes and how to make better use of the elicited expert knowledge. Comparisons among such methods are quite uncommon as extant research usually focuses on each technique separately (Webby & O'Connor, 1996), leading Goodwin (2002) to call for more direct comparisons. Exploring the performance of judgmental integration methods is important for both the efficient design of Forecast Support Systems (FSS) as well as for understanding the conditions for effective elicitation and use of the expert knowledge necessary to improve the functioning of these systems. For instance, the credibility of FSS-generated forecasts might affect expert forecaster's behavior, while frequent disuse of system advice may lead to poor performance in judgmental forecasting (Alvarado-Valencia & Barrero, 2014; Goodwin & Fildes, 1999). Also, the timing of expert intervention may be critically important since not all judgmental adjustments contribute

equally to accuracy (Trapero, Pedregal, Fildes, & Kourentzes, 2013). That is, expert adjustments over FSS forecasts may not always be advantageous and the particular benefits may be a function of when and how the expert judgment is integrated into the forecasting process.

Focusing on the above issues, this paper reports a field experiment that systematically compares three methods for integrating expert judgment with system-generated forecasts. In addition, formal mechanisms to assess the relative expertise of forecasters and the relative credibility of system forecasts are evaluated in companies under real settings. Finally, instances when corrections are needed (i.e., when system forecasts have low accuracy and there is room for improvement) are compared with instances when corrections are not needed (i.e., when adjustments have greater potential to deteriorate accuracy of system-generated predictions because there is small room for accuracy improvement).

2. Literature review and research hypotheses

2.1. Comparison of integration methods

Judgmental integration methods are pervasive, particularly in supply chains where a large number of demand forecasts must be performed to reduce inventory costs and achieve better service levels (Syntetos, Boylan, & Disney, 2009). Companies' operations can benefit from the integration of computer-based forecasting methods with the wider organizational context, where judgment plays an important role (Fildes, Nikolopoulos, Crone, & Syntetos, 2008).

A typical approach to judgmental integration is first to set an automatic baseline (produced by a system using statistical forecast procedures based on historic data) and then to

judgmentally modify these initial forecasts to incorporate contextual knowledge, a process referred to as *Judgmental Adjustment* (Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009). In *Judgmental Adjustment*, the forecaster is usually given the historical time series (in a table, a graph or both formats) and the system forecast and is asked to produce a final forecast.

Judgmental Adjustment may improve accuracy particularly when expert judgments incorporate special events and contextual knowledge into unstable series (Fildes et al., 2009; Goodwin, 2002; Webby & O'Connor, 1996). However, such adjustments might be influenced by several biases, including overconfidence in own judgment (Friedman et al., 2001; Lawrence et al., 2006; Lim & O'Connor, 1996; Sanders, 1997); anchoring and adjustment (i.e., anchoring the forecast in a single cue like the last point or the system forecast, and then making insufficient adjustments to this cue) (Epley & Gilovich, 2006; Fildes et al., 2009; Goodwin, 2005; Lawrence & O'Connor, 1995); and predisposition to adjust (forecasters making too many small harming adjustments to system forecasts without specific reason, leading to deteriorated accuracy) (Fildes et al., 2009; Lawrence et al., 2006; Sanders & Manrodt, 1994; Önkal, Gönül, & Lawrence, 2008). Usually, large and negative adjustments tend to perform better because they show less bias than positive adjustments (Fildes et al., 2009).

In addition to *Judgmental Adjustment*, several integration methods have been proposed as alternatives in the literature. The combination method consists of a simple mathematical aggregation of human and system forecasts. Typically, this combination is a simple average (hereafter called *50-50 Combination*) that has been shown to be robust in several contexts (Blattberg & Hoch, 1990; Franses & Legerstee, 2011). In this method, the forecaster is

usually given the historical time series of the product (in a table, a graph or both formats) and is asked to produce a final forecast. Typically, the forecaster does not know that his/her forecast is going to be combined with a system forecast. Combination has been shown to perform well when inputs are based on independent information sets (Goodwin, 2000; 2002), but the same cognitive biases present in *Judgmental Adjustment* may always appear.

Finally, the *Divide-and-Conquer* method is based on the notion that the system forecast is already based on historical information. Therefore, forecasters should avoid re-assessing historical information because that would lead to inefficient overweighting of past data. The *Divide-and-Conquer* method restricts/prevents human access to this previously computer-modeled information (i.e., the forecaster is not given the time series and the system forecast, but is told how this system forecast is generated), and asks the forecaster whether s/he would like to modify the system forecast (in light of additional information possessed by the forecaster) and, if so, by which amount.

In *Divide-and-Conquer*, decision makers delegate to the system the modeling process of available structured information while focusing their efforts on unmodeled important information that can lead to changing the system advice (Jones & Brown, 2002). Consequently, biases such as anchoring and adjustment might be reduced. However, the lack of information availability might override this advantage. Although this method has been suggested for forecasting tasks (Jones, Wheeler, Appan, & Saleem, 2006; Wright, Saunders, & Ayton, 1988), its applicability has not been tested for the specific case of demand forecasting.

Comparisons of expert elicitation methods have found that advantages of specific methods may be task-dependent; i.e., a direct comparison of elicitation methods on different problems showed that no single approach performed consistently better across all tasks (Flandoli, Giorgi, Aspinall, & Neri, 2011). For instance, conjoint analysis might be preferred when the task is framed as comparisons, while probability elicitation may perform better when a different task structure is used (Dalton, Brothers, Walsh, White, & Whitney, 2013). To the best of our knowledge, very few studies have attempted direct comparisons among judgmental integration methods using real experts in a demand forecasting task. In particular, a formal comparison of *Judgmental Adjustment* and *Divide-and-Conquer* showed that (a) providing the statistical baseline for a judgmental adjustment can lead to more weight given to statistical information as the forecaster tries to incorporate contextual and historical information simultaneously, while (b) encouraging the divide-and-conquer strategy leads to better performance (Jones et al., 2006). Along similar lines, Franses and Legerstee (2013) have demonstrated that formally incorporating judgment may prove helpful when model performance is poor. In an extensive demand forecasting study, Fildes et al (2009) have shown that *50-50 Combination* (also known as the Blattberg-Hoch method) improves accuracy by decreasing the harmful impact of unjustified large (and usually positive) adjustments. Similar results are echoed using non-expert participants (e.g., in extrapolation tasks without contextual information (Webby & O'Connor, 1996)). The current study aims to fill this research gap by focusing on a formal comparison of these three integration methods via a demand forecasting task with real experts in their naturalistic settings. It should be noted that, although group integration methods such as Delphi have been shown to improve forecast accuracy (Armstrong, 2006; Rowe & Wright, 2001), our focus is on individual judgmental integration methods that allow us to isolate the

effects of individual expertise and credibility of system forecasts within an expert knowledge elicitation framework.

The three aforementioned methods exemplify a trade-off between information availability and well-known cognitive biases at an individual level. It may be argued that *Judgmental Adjustment* provides the most information of the three approaches, but may be more subject to the anchor and adjustment bias, precisely due to the amount of information available. *50-50 Combination* withholds a piece of information from the forecaster (i.e., the system forecast), and the forecaster is not allowed to perform the final integration so as to reduce biases. Finally, *Divide-and-Conquer* tries to prevent the forecaster from having two biases: anchoring in past demand/system forecasts and making unjustified adjustments to the system forecast; but carries the associated cost of significantly less available information.

Previous work has shown that debiasing via restricting information access is a hard task (Goodwin, Fildes, Lawrence, & Stephens, 2011). On the other hand, access to relevant information, particularly when it comes from different/independent sources, can improve accuracy if it is well integrated by the methods or by the judges (Bolger & Wright, 2011; Goodwin, 2002; Van Bruggen, Spann, Lilien, & Skiera, 2010). Therefore, we hypothesize:

H1: *Judgmental adjustment will yield the highest accuracy improvement in demand forecasts among the evaluated methods.*

However, our expectation is that there might be some measurable effect on debiasing from the *Divide-and-Conquer* method. In particular, we expect that there would be less anchoring over the system forecast when correction is needed (because the anchor value is not provided) and a lower number of adjustments made when a correction over the system forecast is not needed (by focusing the forecasters initially on having or not a rationale for

the adjustment). Note that this comparison needs to be made against the *Judgmental Adjustment* method because the *50-50 combination* method does not provide system forecasts to the experts (thus banning any corrections to such forecasts).

Accordingly, our hypotheses are as follows:

H2: *The "Divide-and-Conquer" method will lead to less frequent adjustments than the "Judgmental Adjustment" method when correction over the system forecast is not needed.*

H3: *The "Divide-and-Conquer" method will lead to larger adjustments than the "Judgmental Adjustment" method when correction over the system forecast is needed.*

2.2. Expertise and credibility of system forecasts

When integration methods are used in demand forecasting, the resulting forecasts might be affected by the individual's expertise as well as by the perceived credibility of system forecast suggestions (Alvarado-Valencia & Barrero, 2014; Lawrence et al., 2006).

The importance of expertise demands an adequate definition and a measurement of this critical construct. Definitions found in the literature usually refer to at least three components of expertise: first, a field of specialized knowledge where expertise is observable (domain knowledge); second, an outstanding expert's performance in this field; and third, the consistency (i.e., time-lasting and reproducibility) of such performance. Measurement of expertise is usually conducted through comparisons (novice vs expert), peer recognition or objective measures of efficiency, and effectiveness in domain knowledge (Charness & Tuffiash, 2008; Germain & Tejeda, 2012). In expert elicitation, *a priori* selection (based on publication record, group membership or résumé), co-nomination and peer suggestions are frequent (Butler, Thomas, & Pintar, 2015; EPA, 2011; Meyer &

Booker, 2001; Nedeva, Georghiou, Loveridge, & Cameron, 1996) because it is quite difficult to develop tailored tests for knowledge domain effectiveness.

Expertise in the demand-forecasting domain knowledge has been related primarily to intimate product knowledge (Lawrence et al., 2006). This intimate product knowledge allows the expert to be in contact with environmental information that is not captured by statistical models, such as special promotions (Trapero et al., 2013), sudden and unexpected changes in the market, competitors' behavior, and supply-related constraints (Lee, Goodwin, Fildes, Nikolopoulos, & Lawrence, 2007; Webby, O'Connor, & Edmundson, 2005). Therefore, experts in demand forecasting might be found in job positions that are in permanent contact with such unmodeled environmental information.

However, information access is not enough. In addition, it is necessary to have (i) the ability to integrate this information into the final forecast, and (ii) the motivation to do such integration (Gavrilova & Andreeva, 2012). The review of Webby & O'Connor (1996) showed that experiential knowledge of cause-effect relationships encountered in the industry may not be a good predictor of superior accuracy. Another study (Edmundson, Lawrence, & O'Connor, 1988) showed that intimate domain knowledge elicited from experts was useful only for the most important products, but not for the others.

In sales and operations areas, there are several positions where an important part of work performance is to assess or forecast demand formally or informally based on contextual information. For instance, supply chain managers make decisions about when and how much to order for different products. Thus, success in these positions likely depends on the correct assessment of future demand based on information about product rotation and the

possible size of the order of key clients. Marketing and sales managers are expected to take actions to modify the demand and counteract competitors, which requires the ability to correctly foresee the effects of their actions.

If sales and operations experts have different relative expertise contingent on their job positions, and their job performance is related to adequate forecasting, it follows that their job expertise is partially related to an ability to integrate information into forecasts. This ability would be particularly useful when system forecast lacks this information and, as a consequence, a correction over the system forecast is needed. As a result, our fourth hypothesis is:

H4: Higher employee expertise will improve accuracy when correction over the system forecast is needed.

Notice that we expect to verify H4 only if job expertise directly or indirectly requires an assessment of future demand, as explained before, and only where a correction over the system forecast is really needed.

In *Judgmental Adjustment* and *Divide-and-Conquer*, experts relate their expertise to system forecasts' advice. Although it is expected that experts exhibit overconfidence in their own judgment and, as a result, discount the advice (Bonaccio & Dalal, 2006), it can also be expected that individuals show different levels of advice discounting due to different levels of source credibility.

Source credibility is related to a general valuation of the trustworthiness of the trustee outside of the context of specific advice or suggestion (Mayer, Davis, & Schoorman, 1995), and results from a combination of prior information that might be grounded on source reputation, second-hand information on past performance, current experience with the

source, as well as organizational and contextual factors (Alvarado-Valencia & Barrero, 2014). Extant work suggests that credibility of human sources may be assessed differently than credibility of expert systems. Expert systems are believed to be more consistent and less prone to biases than humans. However, expert systems are expected to be less adaptable than human sources, and thus are perceived as being unable to capture all aspects of reality. Expert systems also raise higher performance expectations than humans; as a consequence, expert system errors affect their credibility more severely than human errors (Madhavan & Wiegmann, 2007; Sundar & Nass, 2000). In expert systems literature, the same advice is discounted less when is believed to come from a human expert rather than from an expert system (Lerch, Prietula, & Kulik, 1997; Waern & Ramberg, 1996). Similar results have been found in the judgmental forecasting literature (Onkal, Goodwin, Thomson, Gonul, & Pollock, 2009; Önkal et al., 2008).

At least three mechanisms of the influence of source credibility on discounting system forecast advice are plausible. First at all, research has shown that source credibility is an important factor for persuasion power (Pornpitakpan, 2004), and higher persuasion power may lead to less advice discounting. Second, source credibility is one of the constituents of trust, and higher levels of trust in systems' advice have been found to reduce advice discounting (Goodwin, Gonul, & Onkal, 2013). Finally, if advisors feel that they are relatively less task-expert than the expert system, then less advice discounting could be expected (Rieh & Danielson, 2007).

Therefore, we constructed the following hypothesis:

H5: In Divide-and-Conquer" and "Judgmental Adjustment" methods, there will be an interaction between credibility of system forecasts and forecaster expertise such that:

H5a: larger adjustments will be made when higher forecaster expertise is accompanied by low credibility of system forecasts.

H5b: more frequent adjustments will be made when higher forecaster expertise is accompanied by low credibility of system forecasts.

3. Methods

We conducted a longitudinal field study designed to assess differences in accuracy improvement among three human-computer integration methods: *Judgmental Adjustment*, *50-50 Combination* and *Divide-and-Conquer*. The study assessed the relative expertise and credibility of system forecasts of participants and compared instances where correction of the system forecast was needed to instances when correction was not needed.

3.1. Sample Selection and Characteristics

Companies: Four companies provided access and consent for this field study. We required that the companies were large enough to have at least three different products and were willing to participate in the study. In each company, a key contact person provided assistance with logistics, the selection of products, and the identification of potential participants to be included in the study. This contact person was not included as a participant in the study. The participating companies belonged to different industrial sectors (Table 1). Companies A & C were branches of large multinationals; Companies B & D were local companies, with sales of around US$ 100 million and US$2 million respectively.

Table 1- Participants, products and data-point distribution among companies

	Company

	A	B	C	D
Sector	Chemical	Technology	Food & beverages	Office products retailer
Aggregation level	Product reference	Product family & client type	Product & client type	Product family
Participants	6	10	9	6
Products	4	5	4	4
Collected forecasts	104	248	91	95
Missed forecasts (drop-outs)	4	10	8	7

Products: To be considered in the study, products needed to comply with the following characteristics: a) The product needed to be important for forecasters in terms of volume or value; b) no new products or products close to being discontinued were considered; c) a historical track availability (of at least two years) with non-zero demand was required); d) each product was forecasted on a monthly basis; and e) for each product, at least three participants with extensive product knowledge were available. Meaningful units and aggregation levels were selected for the products based on consultations with the key contacts in each company (Table 1), as each company might have needed different aggregation levels for decision making (Alvarado Valencia & García Buitrago, 2013). The final numbers of selected products per company were quite similar (Table 1). Final selected products had a historical training track ranging from two to eight years, with coefficients of variation in a broad range from 0.32 to 1.31 (Table 2).

Table 2- Product features and data collection

Company	Series	Training length (months)	*CV*	Fitted method	Residuals *CV*	Participants	Months	Collected forecasts
A	1	29	0.64	Seasonal ES	0.74	3	6	18
A	2	29	0.91	Seasonal ES	1.11	3	6	18
A	3	29	0.81	Winters additive	0.55	3	6	18
A	4	44	0.93	Winters additive	0.82	6	9	50
B	5	44	0.73	Seasonal ES	0.94	6	8	48
B	6	44	0.32	Winters additive	0.90	6	8	48
B	7	44	1.11	Seasonal ES	1.11	6	9	51
B	8	44	1.27	Winters additive	1.02	6	9	51
B	9	44	0.72	Seasonal ES	0.75	6	9	50
C	10	41	1.28	Seasonal ES	0.93	3	5	13
C	11	44	0.81	Simple ES	1.00	6	9	49
C	12	41	0.35	Seasonal ES	0.83	3	5	15
C	13	41	0.42	Seasonal ES	0.99	3	5	14
D	14	82	0.42	Seasonal ES	1.03	4	6	23
D	15	82	1.31	Seasonal ES	0.83	4	6	23
D	16	82	0.46	Winters multiplicative	1.11	5	6	27
D	17	82	0.57	Winters additive	1.15	4	6	22

Participants: Selected participants typically worked within the broad sales & operations area (S&OP). All participants were required to have hands-on experience in their assigned products and to have information on these products that could help evaluate and forecast demand based on contextual information, whether formally or informally (although they may not necessarily have experience in the forecasting function within the company). Potential participants were then contacted by email, and an initial interview was set to explain the purpose, scope and research methods, including their right to drop out of the study at any time. All potential participants accepted our invitation and provided informed

consent prior to starting the data collection. All procedures were approved by the Research and Ethics Committee of the School of Engineering at Pontificia Universidad Javeriana.

Participants then completed a survey that contained instruments to measure expertise (Germain & Tejeda, 2012) and credibility of system forecasts (Meyer, 1988). Details of instruments are explained in Section 3.5. Finally, demographic information was collected and a pilot test with each final participant was performed prior to starting the field study to clarify the procedures of the data-collection session. The number of participants per company ranged from six to ten (Table 1). Participants from Company A exhibited higher average age and experience while participants in Company C showed the least variability in age and experience (Table 3).

Table 3- Participant demographics

	Company			
	A(N=6)	B(N=10)	C(N=9)	D(N=6)
Sector	Chemical	Technology	Food & beverages	Office products retailer
Age (years)	*M*=46.16 *SD*=7.08	*M*=36.80 *SD*=7.99	*M*=29.44 *SD*=3.33	*M*=33 *SD*=6.72
Experience in the company (years)	*M*=9.58 *SD*=7.18	*M*=2.88 *SD*=1.92	*M*=3.14 *SD*=1.75	*M*=3.25 *SD*=3.06
Experience with the product(years)	*M*=20.33 *SD*=7.66	*M*=8.48 *SD*=6.86	*M*=1.57 *SD*=0.93	*M*=5.25 *SD*=5.60
Gender (M-F)	5-1	5-5	5-4	5-1

M=mean; *SD*=standard deviation

3.2 Data collection procedures

Each month, an automatic exponential smoothing model was fitted to each product based on the complete historical track available to generate forecasts and their 95% confidence

intervals for the following month. The model, the forecast and the confidence intervals were produced using the automatic features of the SPSS 20 software, including only exponential smoothing methods. Information about the system forecast fit is presented in table 2, including the automatic fitting method selected by the software and the variation coefficient of the residuals after fitting the selected method, which gives an idea of the residual volatility of the series. With a single exception, all demands were seasonal in nature.

Participants were randomly assigned to one of the three methods following Latin Squares randomization in blocks of three months for each group of three forecasters assigned to each product. Information from at least one time block of three months with a minimum of three forecasters was collected for each product. A single participant might be selected for more than one product in the same company.

Company B accounted for roughly 45% of the collected forecasts, and the remaining forecasts were evenly distributed among companies. Due to vacations, meetings, or participants' lack of time, some forecasts were missing (Table 1). Details of collected forecasts by product are given in Table 2, including number of forecasters, number of months collected, and total forecasts collected by product.

Forecasting collection was performed in the first ten days of the month within another administration office in each company. For each month of the study, the researcher provided instructions from a script (Appendix A) to each participant according to the treatment assigned randomly for the given month. None of the forecasts produced was used for decision making or any other purpose within the company. In all treatments,

participants were encouraged to bring their knowledge about the product into their forecast and to give reasons for their final forecasts after delivering it.

In the *Judgmental Adjustment* and *50-50 Combination* treatments, graphs and tables of historical information were produced using default spreadsheet (Excel) settings to improve external validity, and their layout was kept as similar as possible across treatments and periods to avoid format effects. In *Judgmental Adjustment*, a system forecast was included in graphs and tables. *Divide-and-Conquer* participants did not receive any of these graphs/tables. All graphs and tables were presented on a computer screen. The interviewer registered the final demand forecasts produced and recorded the session audio.

3.3 Independent variables

The main independent variable was the human-computer integration method: *Judgmental adjustment*, *50-50 Combination* and *Divide-and-Conquer*. This variable was collected from the treatment assigned each month to each forecaster for each product.

Credibility of system forecasts was measured with Meyer's scale (Meyer, 1988). Among the indexes that have been developed to assess source credibility, Meyer's scale is one of the most validated and used in newspaper credibility research (Roberts, 2010). Although developed in the context of newspapers, it has been successfully applied to other fields, such as advertising and online information (Choi & Lee, 2007; Greer, 2003; Oyedeji, 2007), showing that the questions of the scale are of a general purpose in source credibility. The scale adapted to the purpose of the present study is presented in Appendix B. The results of Meyer's scale were converted into a binary variable. Participants with scale values from zero to two were classified into low credibility of system forecasts (SF

credibility); participants with scale values from three to five were classified into high SF credibility.

Expertise scores were provided by the key contact person in each company using the Germain and Tejeda scale (Germain & Tejeda, 2012). Because expertise was used in this study as a possible independent variable affecting accuracy, using the accuracy results to determine expertise was not considered appropriate. It appears that there are no scales that are capable of measuring intimate product knowledge expertise in the scientific literature, and there are only a few expertise-measuring methods that can be applied or adapted to different contexts (Kuchinke, 1997). The knowledge subscale of the Germain and Tejeda (2012) general scale of expertise recognition was deemed suitable for this research. This subscale intends to measure an employee's expertise in his job. As explained in Section 2.2, employee expertise in jobs where information integration and foresight are constituents of job performance can serve as a proxy of the ability to improve forecasts. The knowledge expertise subscale has general purpose questions related to an expert's field knowledge in her job from the point of view of a colleague (Appendix C). Therefore, it is an expertise measure based on peer recognition. Reliability for the knowledge subscale was high (α=0.92), and factor structure validity was good (Comparative Fit Index=0.93). It should be noted that due to the recent development of this measure, more general reliability and validity tests in different contexts are still needed.

After collection, expertise scores were normalized within each company to avoid potential key contact biases. Participants with standard values over zero were classified as high-expertise participants, whereas those with standard values below or equal to zero were classified as low-expertise participants.

The need for correction of system forecasts was also an independent variable of interest. If the realized demand value for a product was outside the prediction intervals previously calculated using system forecasts (see Section 3.2.), that particular period was labeled as "correction needed". In contrast, if the actual value was located within the 95% prediction interval, that period was marked as "correction not needed".

Lastly, audiotapes of forecast reasons were categorized independently by two researchers and differences were posteriorly reconciled.

3.5 Dependent variables

Improvement in average percentage error (APE) was used as the dependent variable, defined as follows:

$$APE_t = 100 * \frac{|\ Y_t - F_t|}{Y_t}$$

As a consequence,

$$APE_{IMP} = APE_{system\ forecast} - APE_{integrated\ forecast}$$

where Y_t is the actual observed demand outcome at time t and F_t is the final forecast for the product demand produced previously for time t.

APE is a widespread measure of accuracy used in industry (Mentzer & Kahn, 1995), although it has several weaknesses. We did not use scaled error measures such as average scaled errors (ASE) because all data in this study were positive and greater than zero. Consequently, the advantages of scaled measures were reduced, and APE could be selected based on its simplicity and widespread use (Hyndman, 2006).

Additionally, adjustment response measures were estimated to obtain in-depth information about forecasters' behavior when performing *Judgmental Adjustment* or *Divide-and-Conquer*. The calculated measures were adjustment size in absolute value (scaled by the actual demand) and adjustment direction (positive/negative or none).

3.6 Statistical analyses

A four-way ANOVA was conducted to estimate differences in accuracy improvement in APE. Expertise and SF credibility were between-participant variables, whereas the need for correction and the integration methods were within-participant variables. Bonferroni corrected pairwise comparisons were used for post-hoc comparisons.

For the methods of *Divide-and-Conquer* and *Judgmental Adjustment*, Chi square contingency tables were performed to assess the effects of independent variables in adjustment frequency, and a two-way ANOVA was performed to assess the effect of SF credibility and expertise in the absolute adjustment size (scaled by the actual demand).

Also, Chi Square contingency tables were performed to find relationships between rationale types, forecaster behavior and integration methods.

4. Results

4.1. Descriptive results

The expertise and SF credibility levels were evenly distributed across our 31 participants (Table 4). In addition, the distribution of integration methods was even, but there were four times more points in the "correction not needed" treatment due to the nature of the study, in

which the need for correction was only known one month after the data were collected (Table 5).

Table 4 - Expertise and SF credibility of participants

		SF Credibility	
		Low	High
EXPERTISE	Low	7	7
	High	8	9

Table 5 - Data-point distribution among treatments

		Integration method			
		Judgmental Adjustment	*50-50 Combination*	*Divide-and-Conquer*	Total
Correction needed	No	140	137	137	414
	Yes	42	40	42	124
	Total	182	177	179	

4.2. Accuracy improvement results

Judgmental Adjustment was the method with highest accuracy improvement, supporting H1 ($p < .05$, Table 6). Bonferroni-corrected pairwise comparisons show that *Judgmental Adjustment* was significantly better than the *50-50 Combination* ($p = .045$) and *Divide-and-Conquer* ($p < .01$) for APE improvement (Table 7). However, significantly higher-level interactions suggest that the accuracy of *Judgmental Adjustment* should be qualified based on different levels of expertise, SF credibility and the need for correction (Table 6).

*Table 6 - APE improvement related to treatments **

	APE improvement	
Source	F	p-value
Main effects		
Need for correction	26.778	< .001
Expertise	5.381	.021
SF Credibility	2.584	.109
Integration method	6.187	.002
2-way Interaction terms		
Need for correction * expertise	5.603	.018
Need for correction * SF credibility	1.428	.233
Need for correction * integration method	3.723	.025
Expertise * SF credibility	5.483	.020
Expertise * integration method	2.450	.087
SF Credibility * integration method	1.467	.232
3-way Interaction terms		
Need for correction * expertise * SF credibility	3.283	.071
Need for correction * expertise * integration method	3.911	.021
Need for correction * SF credibility * integration method	1.854	.158
Expertise * SF credibility * integration method	7.429	< .001
4-way interaction term		
Need for correction * expertise * SF credibility * integration method	9.269	< .001

*Grey values are significant at the 5% level.

Table 7 - Differences in APE improvement by method

(I) Method	(J) Method	Estimated mean difference (I-J)	Standard Error	p-value	95% confidence interval	
					Lower Bound	Upper bound
Judgmental Adjustment	*50-50 Combination*	.11	.045	.045	.002	.218
Judgmental Adjustment	*Divide-and-Conquer*	.149	.043	.002	.045	.253
50-50 Combination	*Divide-and-Conquer*	.039	.043	N.S.	-.065	.143

Expertise interaction with the need for correction was significant for APE improvement ($p = 0.018$), supporting H4 (Table 6). Higher-level interactions also suggest that this interaction should be qualified on different levels of SF credibility and integration methods.

The fourth-level interaction indicated that the *Judgmental Adjustment* method was superior when high expertise and low SF credibility were present, and correction was needed (Table 8). Its effect size was large enough to mark a difference for expertise and methods in the main effects, and it was the only treatment whose confidence interval for APE improvement was clearly above zero. Therefore, it was the only combination of factors that clearly overcame the system forecast, adding value to the final forecast (Figure 1). The effect was smaller for improvement in the median of the APEs but still held (Table 8), indicating that the results were robust, although their effect size was reduced when the effect of extreme improvements (or deteriorations) was removed. When separate ANOVAs were run on the no-need-for-correction and need-for-correction data, the former did not generate any significant effect on the rest of the independent variables, whereas the latter showed a significant effect in the three-way interaction ($p < .01$).

Regarding the three-way interaction when correction is needed, it is important to highlight that expertise and SF credibility did not generate significant differences in accuracy improvement for the *50-50 Combination* and *Divide-and-Conquer* methods. However, in the *Judgmental Adjustment* method, a specific combination of low SF credibility and high expertise generated improvements in APE that surpassed by more than 80% all other treatment combinations that estimated APE improvement (Figure 1).

Given that the use of APE in averages (as done in ANOVA) might raise concerns about the known biases of this measure and due to the presence of high positive skewness and high kurtosis in our APE improvement results, we repeated the analysis using the median of APEs as the estimator (MedAPE) and using average scaled errors (ASE) as the accuracy measure. Findings on significance remained the same, although the size effects changed.

To reduce the chance of spurious p-values and concerns about sample size, we conducted two additional analyses. First, a four-fold random cross-validation showed that results were robust; a time-based analysis of the data showed that relevant p-values appeared when approximately half the sample was collected and did not oscillate between significance and non-significance after two thirds of the sample was collected, supporting convergence over the current sample size.

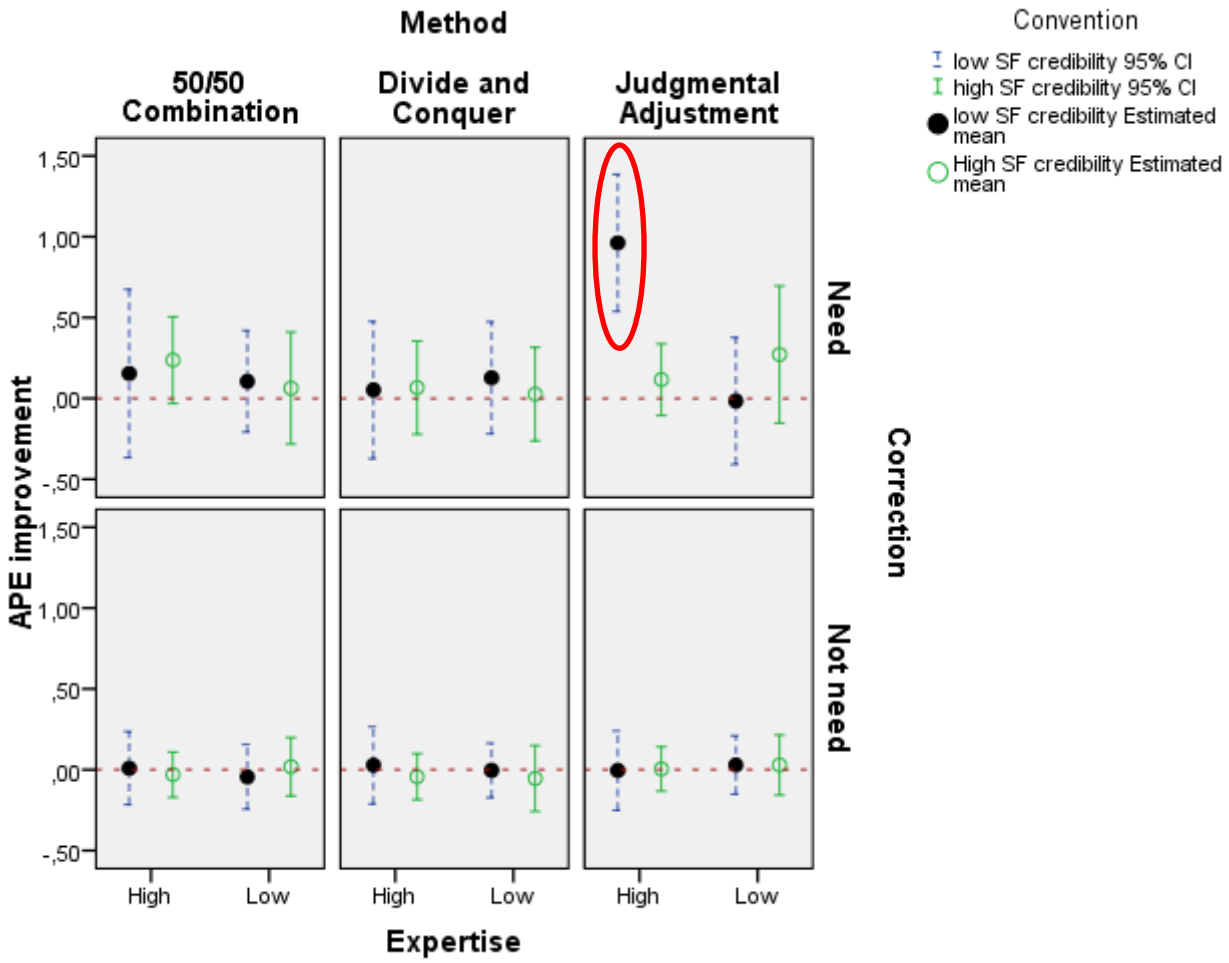
Table 8 - Estimated means and medians of errors for field study treatments

Need for Correction	Expertise	SF Credibility	Method	Estimated MAPEimp	Estimated MdAPEimp	Standard error	99,9% confidence interval	
							Lower bound	Upper bound
No	Low	Low	Judgmental	,029	,000	,055	-,152	,210
			Combination	-,045	,004	,060	-,245	,156
			Divide	-,004	,000	,051	-,173	,164
		High	Judgmental	,029	,000	,056	-,158	,215
			Combination	,019	-,005	,055	-,162	,200
			Divide	-,054	,000	,062	-,258	,149
	High	Low	Judgmental	-,005	,011	,074	-,250	,240
			Combination	,009	,041	,069	-,217	,236
			Divide	,028	,001	,072	-,211	,266
		High	Judgmental	,005	,000	,042	-,133	,142
			Combination	-,030	-,012	,042	-,171	,110
			Divide	-,043	,000	,043	-,186	,100
Yes	Low	Low	Judgmental	-,016	,000	,119	-,409	,377
			Combination	,105	,027	,095	-,208	,419
			Divide	,127	,000	,105	-,219	,474
		High	Judgmental	,272	,211	,128	-,152	,696
			Combination	,063	,000	,105	-,283	,410
			Divide	,026	,000	,087	-,262	,315
	High	Low	Judgmental	,962	,338	,128	,537	1,386
			Combination	,154	,095	,157	-,366	,674
			Divide	,052	,086	,128	-,373	,476
		High	Judgmental	,116	,070	,067	-,106	,337
			Combination	,237	,143	,081	-,032	,505
			Divide	,067	,000	,087	-,222	,355

Table conventions:
Judgmental= *Judgmental Adjustment*
Combination= *50-50 Combination*
Divide= *Divide-and-Conquer*

Figure 1 - Estimated means and 95% confidence intervals for SF credibility, expertise, need for correction and integration method

The circle shows the only treatment with a significant APE improvement over zero.



4.3. Adjustment behavior results

The method was not related to the adjustment frequency when a correction was not needed ($p = 0.433$) or to the adjustment size when a correction was needed ($p = 0.567$); therefore, there is no support for H3 or H2.

The interaction between expertise and SF credibility for absolute size ($p = 0.016$) and adjustment frequency direction relations were found to be significant ($p < 0.001$ for expertise in low SF credibility, but $p > 0.005$ for expertise in high SF credibility). As shown in Figure 2, participants in the high-expertise/low-SF credibility condition tended to make more positive adjustments (standardized residual=2.5), whereas participants in the low expertise condition with low SF credibility conditions avoided making changes to suggestions (standardized residual=2.4), supporting H5b. Although they made less frequent adjustments than expected, participants with low expertise and low SF credibility tended to perform adjustments of a larger size, as shown in Figure 3, yielding the result contrary to the one expected for H5a.

The best adjustments were negative ones when they were really needed, 95% CI [0.30, 0.49] as shown in Figure 4. However, when correction was not needed, negative adjustments also significantly improved accuracy over positive adjustments, 95% CI [0.04, 0.13].

Figure 2 - Percentage of adjustment type occurrence by groups of expertise and SF credibility

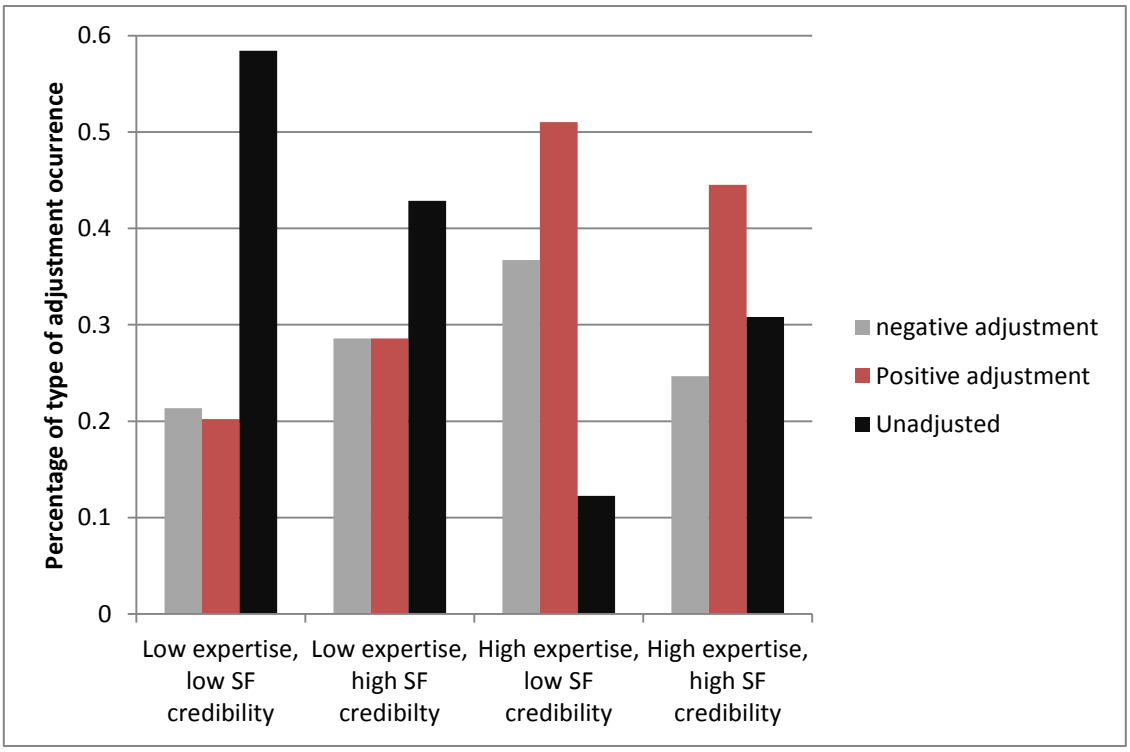


Figure 3 - Adjustment size by SF credibility and expertise (*Divide-and-Conquer* and *Judgmental Adjustment* treatments only)

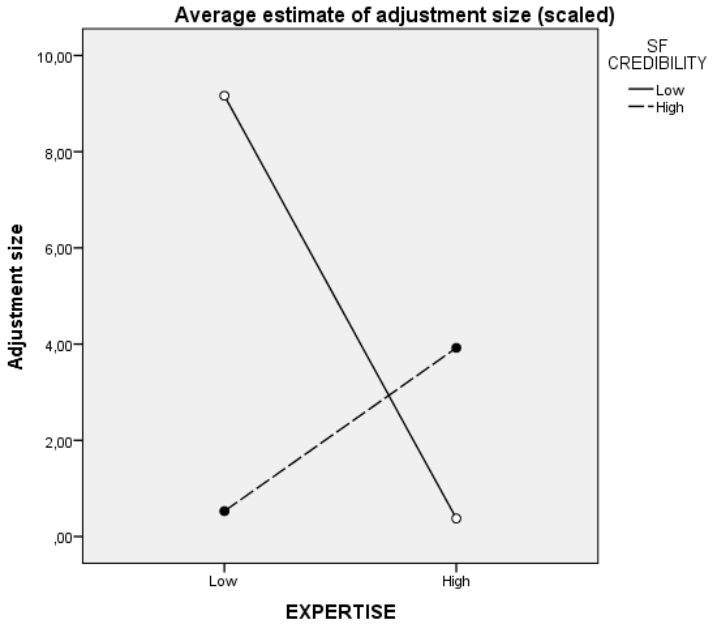
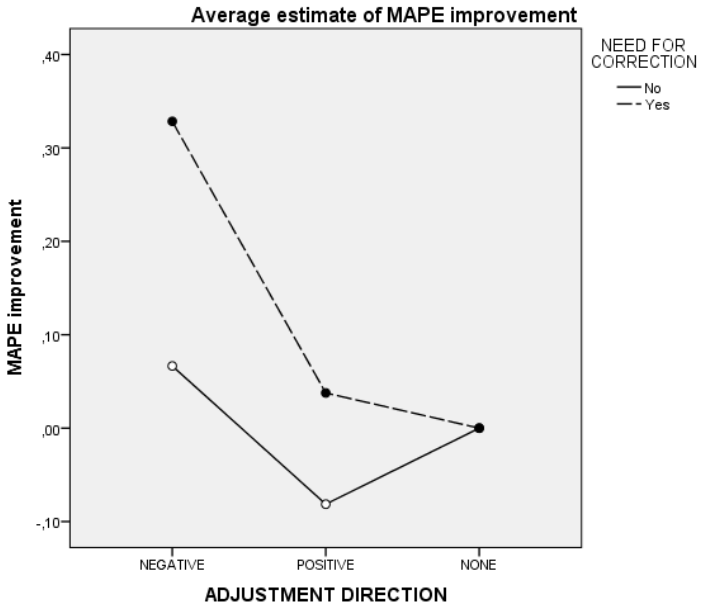


Figure 4 - MAPE improvement by adjustment direction vs. need for correction



4.4. Expert rationales

Experts did not give reasons for their final forecasts in 28.6% of the collected forecasts. A single causal force to justify the final forecast was elicited in 62.4% of the occasions, and multiple causal forces were elicited in only 9% of the cases. When rationales where provided, its average length was 29 words with a high variability (ranging from 1 to 260 words).

Five main types of reasons were elicited from experts when asked for rationales to produce their final forecasts in all companies. First, historical reasons were quoted (40.4%). These reasons included seasonality on an annual basis and long-term trends of ascent or descent. In the *50-50 Combination* method, these reasons were usually further elaborated comparing specific figures, whereas in the other methods, historical reasons were shorter and more direct.

Second, marketing actions were mentioned (25.5%). Beyond specific promotions, there were also advertising plans, brand awareness strategies, new strategic deals close to being sealed, and new distribution strategies that made experts believe that the forecast should be changed.

Third, supply chain reasons were given specific importance in judgmental forecasting (17.7%). Reasons included both sides of the supply chain. The main reasons were current inventory levels, whether in the distributor or in the company, but previously settled pre-orders and lead times were also cited.

Fourth, reasons related to organizational goals and job performance were cited (9.4%). These reasons to settle a forecast included quarter or end-of-year deadlines and job responsibilities to comply with target sales, linked to perceived control.

Finally, business environment reasons out of the company control were mentioned, including economy and market trends, legal decisions, and competition actions (7%).

There were also reasons related to the specific industry and sector. In the technology company, product life-cycle reasons were also frequently cited, whereas in the chemical company, the weather was mentioned as an important causal force. These were included in the previous analysis under appropriate labels.

Rationales were found to be significantly related to the integration method used ($p < .01$). In *50-50 Combination,* historical reasons were quoted more than expected, whereas in *Divide-and-Conquer* historical reasons were quoted less than expected. There was also a relationship between type of rationales and adjustment direction ($p < .01$). Negative adjustments were more frequent on supply and business environment reasons, whereas positive adjustments were more frequent when marketing and organizational goal rationales were quoted.

5. Discussion

This study aimed to compare the accuracy improvement and adjustment behavior of three human-computer integration methods used to generate demand forecasts with real products and practitioners. The study considered the potential effects of three important variables on

the resulting accuracy of the forecasts and the behavior of forecasters, i.e., the need for the correction of the system forecast, the relative expertise of the forecaster and the relative credibility of the forecaster in the system forecast. We observed improvement in the accuracy for forecasters with higher relative expertise and low credibility of system forecasts when correction over the system forecast was needed and the *Judgmental Adjustment* method was applied. We also observed different adjustment behavior patterns related to different levels of expertise and credibility of system forecasts.

Our study proposed and tested the use of a general scale of employee job expertise as a proxy to discriminate between levels of demand forecasting domain knowledge. Although we strongly encourage further studies to test the validity and reliability of this scale, the preliminary results are promising. This scale is based on peer ratings and therefore is subject to power biases in organizations, but it is clearly an improvement over widely used low structured methods to select experts from public recognition or co-nomination. It can be combined with the development of short questionnaires tailored to the specific domain knowledge and with measures of personal characteristics recently found to improve forecasting in other fields (Mellers et al., 2015). However, our results show that the development of a more structured mechanism to discriminate expertise in demand forecasting judgmental adjustments is a task worth attempting.

Although it is reasonable to expect higher accuracy among employees who had more expertise, it is also true that research so far has found mixed evidence (Lawrence et al., 2006). These mixed results might be due to the hidden interaction effects of other variables not considered simultaneously in previous studies. In this research, we studied three variables (i.e., credibility of system forecasts, integration method and need for correction)

that may explain why experts sometimes do perform better and sometimes do not. We found that experts perform well particularly when perceived credibility of system forecasts is low. A possible explanation is that low levels of confidence in system suggestions allow experts to detach themselves enough from the system forecast when really needed, avoiding the anchor and adjustment heuristic. This explanation suggests that healthy skepticism about system forecasts may reduce the possible anchoring effects of such forecasts, and such bias reduction becomes particularly useful when experts believe that they need to modify the system forecast by a substantial amount. A possible subsequent laboratory experiment could evaluate the strength of the anchor and adjustment heuristic with different levels of credibility of system forecasts, presenting simultaneously the system forecast with rationales to substantially modify it.

A second explanation is that low credibility reduced complacency with support systems suggestions (Goddard, Roudsari, & Wyatt, 2012) and therefore motivated forecasters to add their knowledge and feel accountable for the results (Fildes, Goodwin, & Lawrence, 2006; Wright, Lawrence, & Collopy, 1996). In support of that, forecasters with relatively high expertise tended to make adjustments in almost all the cases. This tendency may occur because they think they need to contribute somehow to the forecast (Gonul, Onkal, & Goodwin, 2009). In our study, this pattern of highly frequent adjustment increased if perceived credibility of system forecasts was low.

Additionally, the benefit in accuracy of having an expert was observed only when the *Judgmental Adjustment* method of integration was used. A possible explanation is that *Judgmental Adjustment* was the only method in this study that allowed the forecaster to access all relevant information, and none of the other methods reduced biases significantly

to overcome this information loss. The *50-50 combination* method made the forecasters focus on detecting historical trends (as revealed in their forecast rationales), thus effectively underweighting the additional contextual knowledge they may have. In this way, system and forecaster inputs were not independent; consequently, *50-50 Combination* underperformed against *Judgmental Adjustment*. In contrast, in the *Divide-and-Conquer* method, a lack of access to system forecasts made it difficult for experts to assess the quality of system advice as well as the amount of correction needed. The *Divide-and-Conquer* method did not show that it reduced the frequency of adjustment when correction was not needed, and it did not increase the adjustment size when correction was needed. The overall implication of our results is that trying to debias forecasters through information restrictions did not work, whereas providing access to all relevant information to experts was helpful to assess the need for change. Additionally, information access may offer more control to the forecaster, which may result in a sense of satisfaction or comfort while doing the task. A follow-up study can test if providing only the system forecast would be enough to give access to all relevant information, because the system forecast can be regarded as a summary of the historical track.

The judgmental integration task can be regarded as a joint effort between support systems and experts to develop a better forecast. In this regard, an analysis of process gains versus losses of expert knowledge elicitation can be conducted (Bedard, Biggs, Maroney, & Johnson, 1998; Rowe, Wright, & Bolger, 1991). *Divide-and-Conquer* was unable to deliver process gains through bias reduction, and possibly generated process losses by forbidding access to relevant information to participants. *50-50 combination* generated expertise overlap by focusing the expert on the same information that the support system is already

assessing and reducing the chances of the inclusion of diverse inputs and knowledge into the task. As a consequence, this procedure did not generate process gains through an integration that surpasses the sum of the parts. Meanwhile, *Judgmental Adjustment* was closer to a group process where an expert is faced with another suggestion (the system advice) and can potentially generate a process gain through knowledge pooling and sharing. However, the presence of possible biases such as anchor and adjustment require the presence of healthy skepticism from experts in order to avoid process losses.

Finally, our results indicate that when corrections were not needed, negative adjustments improved accuracy whereas positive adjustments deteriorated it, leading to a net sum of no improvement. Therefore, expertise still contributed in occasions when small adjustments (i.e., adjustments when the realized value falls inside the 95% interval of the system forecast) were required, but its effects were obscured by overoptimism and predisposition to adjust. When corrections were needed, the overall result of adjustments was an improvement of accuracy, due to the benefits of well-sized adjustments that were usually negative.

We observed associations between negative adjustments and business environment and supply chain reasons in the studied companies. Supply chain and business environment reasons can reflect a current state of affairs outside company control (while clearly affecting possible demand outcomes), whereas market actions and goal-oriented reasons may be related more to a bet into the future, partially depending on business actions and being mediated by illusion of control) . These results led us to suggest emphasizing elicitation of knowledge on situations outside company control over the elicitation of

knowledge related to plans or promotions under the company control that can be modeled through statistical analysis (Trapero et al., 2013).

At least three limitations of the study should be discussed. First, participants changed methods randomly every month and were not given feedback on their performance; therefore, there is no way of evaluating possible learning effects. However, a previous study found small or no learning effects with outcome feedback (Lim & O'Connor, 1995). Second, the elicited forecasts had no consequences for the company decision making or the performance evaluation of participants in companies; therefore, we do not expect any major effects on their responses related to the political pressures or organizational cultures. Although the nature of our task did not allow the complete appearance of such effects, the presence of adjustment reasons related to goals and perceived control is an indication that such pressures played at least a partial role in our task. We believe that the implication of our results will be valid in real settings.

Finally, the study included a set of specific companies and selected products. Differences among industries, although not evaluated in this study, can be an avenue for future research. In addition, participants in the study were asked to focus on a few products, whereas in real settings, forecasters typically are required to forecast a huge number of products in specific locations. The consequences are twofold. First, our belief is that intimate product knowledge—and therefore judgmental adjustment expertise—is practically impossible for every disaggregation level in such a large task with the usual time restrictions. We selected experts and aggregation levels for each product where environmental and product knowledge can be elicited, and the same needs to be done to apply our results in real settings. A possible future research direction could entail best

practices in selecting those products and aggregation levels where expertise can clearly make a difference. Second, given the aggregation level, intermittent demand was not considered. The next step would be to evaluate the possible extensions of the present study to explore this important forecasting problem.

Although it is true that generalizations should not be made outside this context, we believe our efforts to work with companies in different sectors, with participants having a wide range of demographic characteristics and experience, in physical settings that resemble the day-to-day conditions of forecasters in their workplace, and with products that are important to their positions within their companies provide a desirable backdrop in forecasting expertise and related processes. Therefore, our results may be viewed as providing a robust starting point to exploring expertise and credibility of system forecasts issues for forecasters in similar industrial sectors.

6. Implications for forecasting practice

Using experts in a demand forecasting field study, current work addresses important issues of expert knowledge elicitation in a real (and ecologically valid) forecasting framework. Our findings contribute to the discussion of four important questions widely discussed in forecasting practice.

The first question relates to the conditions for forecaster selection when adjusting system forecasts. Our results show that expertise in key job positions is a necessary but not sufficient condition for good forecasting performance. In order to be able to modify the

system forecast in the right amount when it is really needed, such expertise needs to be combined with healthy skepticism about credibility of system forecasts.

Second, should companies invest to improve adjustment processes or is it better (for accuracy) to rely on system projections? Our results suggest that, in the long run and with proficient selection of experts, accuracy is improved with *Judgmental Adjustment* when useful information is incorporated into the adjustment process. If a greater gain is sought, group processes effectively managed to elude political and organizational pressures and integrate individual adjustments (such as Delphi) will definitely prove valuable. Proposed categorization of adjustment rationales developed in this study might be a good starting point for scenario construction to aid with such processes, encouraging particularly the analysis of causal forces outside business control.

Another possible mechanism for greater gain relates to bias-reduction techniques for expert forecasters. Restricting information to experts does not appear to be a desirable method to avoid biases. Training to foster a healthy skepticism towards system forecasts can be reinforced by highlighting the limitations of system forecasts in situations like inventory shortages or environmental/ structural changes, emphasizing potential negative effects. Overoptimism could be reduced by challenging attempted positive adjustments (e.g., based on company plans and promotions), while letting negative adjustments go unchallenged. Periodic training and feedback mechanisms to track and combat well-known cognitive biases (e.g., overconfidence, desirability bias) may also prove effective (Benson & Önkal, 1992).

A third question relates to whether system forecasts should be given to forecasters. Despite the possible biases generated by the presence of a system forecast, such as anchoring and excessive weighing of past data, it appears that on average the availability of the system forecast improves the accuracy of the integrated forecasts. Additionally we have observed that the practitioners were quite uneasy when this information was lacking, a phenomenon that may be related to a perceived loss of control and/or reduced confidence emanating from lack of a starting benchmark. Further work comparing presence/absence of system forecasts and/or historical information might help elucidate whether the system forecast can replace the historical information. Along similar lines, research that incorporates qualitative methodologies to study expert knowledge elicitation will prove extremely useful to better understand the reasons behind forecasters' use/misuse of system forecasts.

Lastly, although it is not possible to know in advance if a modification will be needed to improve forecasting performance, there are occasions when such tweaks are clearly warranted: for instance, when there are supply chain restrictions, structural changes in time series, inflection points in product life cycle, sudden environmental changes, and/ or anticipated competitor moves. In these occasions, our findings suggest encouraging systematic use of *Judgmental Adjustment* with forecasters who possess (i) high expertise in the judgmental forecasting domain knowledge, along with (ii) healthy skepticism about support system advice that encourages a realistic/unbiased assessment of system forecasts.

7. References

Alvarado-Valencia, J. A., & García Buitrago, J. A. (2013). Selecting and using an adequate disaggregation level in time series forecasting: A study case in a subscription business

model company through the analytic hierarchy process. *Selección y utilización de niveles de desagregación adecuados en pronósticos de series temporales: Caso de estudio en una empresa de suscripción utilizando el proceso analítico jerárquico, 15*(1), 45-64.

Alvarado-Valencia, J. A., & Barrero, L. H. (2014). Reliance, trust and heuristics in judgmental forecasting. *Computers in human behavior, 36*(0), 102-113. doi:http://dx.doi.org/10.1016/j.chb.2014.03.047

Armstrong, J. S. (2006). Findings from evidence-based forecasting: Methods for reducing forecast error. *International Journal of Forecasting, 22*(3), 583-598.

Aspinall, W. (2010). A route to more tractable expert advice. *Nature, 463*(7279), 294-295.

Bedard, J. C., Biggs, S. F., Maroney, J. J., & Johnson, E. N. (1998). Sources of Process Gain and Loss From Group Interaction in Performance of Analytical Procedures. *Behavioral Research in Accounting, 10*, 207-239.

Benson, P. G., & Önkal, D. (1992). The effects of feedback and training on the performance of probability forecasters. *International Journal of Forecasting, 8*(4), 559-573. doi:10.1016/0169-2070(92)90066-I

Blattberg, R. C., & Hoch, S. J. (1990). Database Models and Managerila Intuition:50% model + 50% manager. *Management Science, 36*(8), 887-899.

Bolger, F., & Rowe, G. (2014). Delphi: Somewhere between Scylla and Charybdis? *Proceedings of the National Academy of Sciences, 111*(41), E4284-E4284. doi:10.1073/pnas.1415425111

Bolger, F., & Rowe, G. (2015). The Aggregation of Expert Judgment: Do Good Things Come to Those Who Weight? *Risk Analysis, 35*(1), 5-11. doi:10.1111/risa.12272

Bolger, F., & Wright, G. (2011). Improving the Delphi process: Lessons from social psychological research. *Technological Forecasting and Social Change, 78*(9), 1500-1513. doi:http://dx.doi.org/10.1016/j.techfore.2011.07.007

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes, 101*(2), 127-151.

Budnitz, R. J., Apostolakis, G., Boore, D. M., Cluff, L. S., Coppersmith, K. J., Cornell, C. A., & Morris, P. A. (1997). *Recommendations for Probabilistic Seismic Hazard Analysis: Guidance on Uncertainty and Use of Experts*. Retrieved from Washington, D.C.:

Butler, A. J., Thomas, M. K., & Pintar, K. D. M. (2015). Systematic review of expert elicitation methods as a tool for source attribution of enteric illness. *Foodborne Pathogens and Disease, 12*(5), 367-382. doi:10.1089/fpd.2014.1844

Charness, N., & Tuffiash, M. (2008). The role of expertise research and human factors in capturing, explaining, and producing superior performance. *Human Factors, 50*(3), 427-432. doi:10.1518/001872008x312206

Choi, S. M., & Lee, W. N. (2007). Understanding the impact of direct-to-consumer (DTC) pharmaceutical advertising on patient-physician interactions - Adding the web to the mix. *Journal of Advertising, 36*(3), 137-149. doi:10.2753/joa0091-3367360311

Cooke, R. (1991). *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford: Oxford University Press.

Dalton, A., Brothers, A., Walsh, S., White, A., & Whitney, P. (2013). Expert elicitation method selection process and method comparison *Neuroscience and the Economics of Decision Making* (pp. 182-194).

Edmundson, B., Lawrence, M., & O'Connor, M. (1988). The Use Of Non-Time Series Information In Sales Forecasting. *Journal of Forecasting, 7*(3), 201.

EPA, U. S. (2011). Expert Elicitation Task Force White Paper. Washington D.C.: Science and Technology Policy Council.

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic. *Psychological Science, 17*(4), 311-318. doi:10.1111/j.1467-9280.2006.01704.x

Fildes, R., Goodwin, P., & Lawrence, M. (2006). The design features of forecasting support systems and their effectiveness. *Decision Support Systems, 42*(1), 351-361.

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. *International Journal of Forecasting, 25*(1), 3-23.

Fildes, R., Nikolopoulos, K., Crone, S. F., & Syntetos, A. A. (2008). Forecasting and operational research: a review. *The Journal of the Operational Research Society, 59*(9), 1150-1172.

Flandoli, F., Giorgi, E., Aspinall, W. P., & Neri, A. (2011). Comparison of a new expert elicitation model with the Classical Model, equal weights and single experts, using a cross-validation technique. *Reliability Engineering and System Safety, 96*(10), 1292-1310. doi:10.1016/j.ress.2011.05.012

Franses, P. H., & Legerstee, R. (2011). Combining SKU-level sales forecasts from models and experts. *Expert Systems with Applications, 38*(3), 2365-2370. doi:10.1016/j.eswa.2010.08.024

Franses, P. H., & Legerstee, R. (2013). Do statistical forecasting models for SKU-level data benefit from including past expert knowledge? *INTERNATIONAL JOURNAL OF FORECASTING, 29*(1), 80-87. doi:10.1016/j.ijforecast.2012.05.008

Friedman, C., Gatti, G., Elstein, A., Franz, T., Murphy, G., & Wolf, F. (2001). Are clinicians correct when they believe they are correct? Implications for medical decision support. *Stud Health Technol Inform, 84*(Pt 1), 454-458.

Gavrilova, T., & Andreeva, T. (2012). Knowledge elicitation techniques in a knowledge management context. *Journal of Knowledge Management, 16*(4), 523-537. doi:10.1108/13673271211246112

Germain, M. L., & Tejeda, M. J. (2012). A preliminary exploration on the measurement of expertise: An initial development of a psychometric scale. *Human Resource Development Quarterly, 23*(2), 203-232. doi:10.1002/hrdq.21134

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association, 19*(1), 121-127. doi:10.1136/amiajnl-2011-000089

Gonul, S., Onkal, D., & Goodwin, P. (2009). Expectations, Use and Judgmental Adjustment of External Financial and Economic Forecasts: An Empirical Investigation. *Journal of Forecasting, 28*(1), 19-37. doi:10.1002/for.1082

Goodwin, P. (2000). Correct or combine? Mechanically integrating judgmental forecasts with statistical methods. *International Journal of Forecasting, 16*(2), 261-275.

Goodwin, P. (2002). Integrating management Judgment and statistical methods to improve short-term forecasts. *Omega- International Journal of Management Science, 30*(2), 127-135.

Goodwin, P. (2005). Providing support for decisions based on time series information under conditions of asymmetric loss. *European Journal of Operational Research, 163*, 388-402.

Goodwin, P., & Fildes, R. (1999). Judgmental forecasts of time series affected by special events: does providing a statistical forecast improve accuracy? *Journal of Behavioral Decision Making, 12*, 37-53.

Goodwin, P., Fildes, R., Lawrence, M., & Stephens, G. (2011). Restrictiveness and guidance in support systems. *Omega, 39*(3), 242-253.

Goodwin, P., Gonul, M. S., & Onkal, D. (2013). Antecedents and effects of trust in forecasting advice. *International Journal of Forecasting, 29*(2), 354-366. doi:10.1016/j.ijforecast.2012.08.001

Greer, J. D. (2003). Evaluating the Credibility of Online Information: A Test of Source and Advertising Influence. *Mass Communication and Society, 6*(1), 11-28. doi:10.1207/s15327825mcs0601_3

Hyndman, R. K., Anne B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*, 679-688.

Jones, D. R., & Brown, D. (2002). The division of labor between human and computer in the presence of decision support system advice. *Decision Support Systems, 33*(4), 375-388. doi:http://dx.doi.org/10.1016/S0167-9236(02)00005-2

Jones, D. R., Wheeler, P., Appan, R., & Saleem, N. (2006). Understanding and attenuating decision bias in the use of model advice and other relevant information. *Decision Support Systems, 42*(3), 1917-1930.

Kuchinke, K. P. (1997). Employee Expertises The Status of the Theory and the Literature. *Performance Improvement Quarterly, 10*(4), 72-86. doi:10.1111/j.1937-8327.1997.tb00068.x

Lawrence, M., Goodwin, P., Oconnor, M., & Onkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting, 22*(3), 493-518.

Lawrence, M., & O'Connor, M. (1995). The Anchor and Adjustment Heuristic in Time-series Forecasting. *Journal of Forecasting, 14*(5), 443-451.

Lee, W., Goodwin, P., Fildes, R., Nikolopoulos, K., & Lawrence, M. (2007). Providing support for the use of analogies in demand forecasting tasks. *International Journal of Forecasting, 23*(3), 377-390.

Lerch, F. J., Prietula, M. J., & Kulik, C. T. (1997). The Turing effect: the nature of trust in expert system advice. In P. J. Feltovich & K. M. Ford (Eds.), *Expertise in Context: Human and Machine* (pp. 417-448). Cambridge, MA: The MIT Press.

Lim, J., & O'Connor, M. (1996). Judgmental forecasting with time series and causal information. *International Journal of Forecasting, 12*(1), 139-153.

Lim, J. S., & O'Connor, M. (1995). Judgemental adjustment of initial forecasts: Its effectiveness and biases. *Journal of Behavioral Decision Making, 8*(3), 149-168.

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human-human and human-automation trust: an integrative review. *Theoretical Issues in Ergonomics Science, 8*(4), 277-301.

Martin, T. G., Burgman, M. A., Fidler, F., Kuhnert, P. M., Low-Choy, S., McBride, M., & Mengersen, K. (2012). Eliciting Expert Knowledge in Conservation Science. *Conservation Biology, 26*(1), 29-38. doi:10.1111/j.1523-1739.2011.01806.x

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709-734.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., . . . Tetlock, P. (2015). Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions. *Perspectives on Psychological Science, 10*(3), 267-281. doi:10.1177/1745691615577794

Mentzer, J. T., & Kahn, K. B. (1995). Forecasting technique familiarity, satisfaction, usage, and application. *Journal of Forecasting, 14*(5), 465-476.

Meyer, M. A., & Booker, J. M. (2001). *Eliciting and Analyzing Expert Judgment: A Practical Guide*: Society for Industrial and Applied Mathematics.

Meyer, P. (1988). Defining and Measuring Credibility of Newspapers: Developing an Index. *Journalism & Mass Communication Quarterly, 65*(3), 567-574.

Morgan, M. G. (2014). Use (and abuse) of expert elicitation in support of decision making for public policy. *Proceedings of the National Academy of Sciences, 111*(20), 7176-7184. doi:10.1073/pnas.1319946111

Nedeva, M., Georghiou, L., Loveridge, D., & Cameron, H. (1996). The use of co-nomination to identify expert participants for Technology Foresight. *R&D Management, 26*(2), 155-168. doi:10.1111/j.1467-9310.1996.tb00939.x

Onkal, D., Goodwin, P., Thomson, M., Gonul, S., & Pollock, A. (2009). The Relative Influence of Advice From Human Experts and Statistical Methods on Forecast Adjustments. *Journal of Behavioral Decision Making, 22*(4), 390-409. doi:10.1002/bdm.637

Önkal, D., Gönül, M. S., & Lawrence, M. (2008). Judgmental Adjustments of Previously Adjusted Forecasts. *Decision Sciences, 39*(2), 213-238.

Oyedeji, T. A. (2007). The Relation Between the Customer-Based Brand Equity of Media Outlets and Their Media Channel Credibility: An Exploratory Study. *International Journal on Media Management, 9*(3), 116-125. doi:10.1080/14241270701521725

Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology, 34*(2), 243-281. doi:10.1111/j.1559-1816.2004.tb02547.x

Rieh, S. Y., & Danielson, D. R. (2007). Credibility: A multidisciplinary framework. *Annual Review of Information Science and Technology, 41*, 307-364. doi:10.1002/aris.2007.1440410114

Roberts, C. (2010). Correlations Among Variables in Message and Messenger Credibility Scales. *American Behavioral Scientist, 54*(1), 43-56. doi:10.1177/0002764210376310

Rowe, G., Wright, G., & Bolger, F. (1991). Delphi: A reevaluation of research and theory. *Technological Forecasting and Social Change, 39*(3), 235-251. doi:http://dx.doi.org/10.1016/0040-1625(91)90039-I

Rowe,G.,& Wright,G.(2001).Expert opinions in forecasting: The role of the Delphi technique. In J.S. Armstrong (Ed.), Principles of forecasting (pp. 125-144). Boston: Kluwer Academic Publishers.

Sanders, N., & Manrodt, K. B. (2003). The efficacy of using judgmental versus quantitative forecasting methods in practice. *Omega, 31*(6), 511-522.

Sanders, N. R. (1997). The Status of Forecasting in Manufacturing Firms. *Production & Inventory Management Journal, 38*(2), 32-36.

Sanders, N. R., & Manrodt, K. B. (1994). Forecasting Practices in US Corporations: Survey Results. *Interfaces, 24*(2), 92-100.

Sundar, S. S., & Nass, C. (2000). Source orientation in human-computer interaction: Programmer, networker, or independent social actor? *Communication Research, 27*(6), 683-703.

Syntetos, A. A., Boylan, J. E., & Disney, S. (2009). Forecasting for inventory planning: a 50-year review. *Journal of The Operational Research Society, 60*, 149-160.

Trapero, J. R., Pedregal, D. J., Fildes, R., & Kourentzes, N. (2013). Analysis of judgmental adjustments in the presence of promotions. *International Journal of Forecasting, 29*(2), 234-243. doi:10.1016/j.ijforecast.2012.10.002

Van Bruggen, G. H., Spann, M., Lilien, G. L., & Skiera, B. (2010). Prediction Markets as institutional forecasting support systems. *Decision Support Systems, 49*(4), 404-416.

Waern, Y., & Ramberg, R. (1996). People's perception of human and computer advice. *Computers in human behavior, 12*(1), 17-27.

Webby, R., & O'Connor, M. (1996). Judgemental and statistical time series forecasting: a review of the literature. *International Journal of Forecasting, 12*(1), 91-118.

Webby, R., O'Connor, M., & Edmundson, B. (2005). Forecasting support systems for the incorporation of event information: An empirical investigation. *International Journal of Forecasting, 21*(3), 411-423.

Wright, G., Lawrence, M. J., & Collopy, F. (1996). The role and validity of judgment in forecasting. *International Journal of Forecasting, 12*(1), 1-8.

Wright, G., Saunders, C., & Ayton, P. (1988). The consistency, coherence and calibration of holistic, decomposed and recomposed judgemental probability forecasts. *Journal of Forecasting, 7*(3), 185-199. doi:10.1002/for.3980070304

Appendix A

50-50 Combination:

This month, we are going to generate a demand forecast for product (name *of the product*) in (*units: dollars, number of items…*). If you don't understand the product definition, please ask for clarification. The screen is showing the historical demand for this product during the last (*number of periods*) periods in the graph and in the table. You are free to consult any additional (non-historical) information you already have that might be related to the product and their business development. Please indicate what you think the demand will be for this product (*name of the product*) in (*units: dollars, number of items…*) for next month, taking into account your judgment and knowledge of the product and the business.

(After the forecast is produced) Please explain your motivations and reasons for this result.

Judgmental Adjustment:

This month, we are going to generate a demand forecast for product (name *of the product*) in (*units: dollars, number of items…*). If you don't understand the product definition, please ask for clarification. The screen is showing the historical demand for this product during the last (*number of periods*) periods in the graph and in the table and a system forecast for the following month in the graph and in the table. This forecast has taken in account three elements: historical trend of data, seasonal effects, and increasing/decreasing effects. You are free to consult any additional (non-historical) information you already have that might be related to the product and their business development. Please indicate what you think the demand will be for this product (name *of the product*) in (*units: dollars, number of items…*) for next month, taking into account your judgment and knowledge of the product and the business.

(*After the forecast is produced*) Please explain your motivations and reasons for this result.

Divide-and-Conquer:

This month, we are going to generate a demand forecast for product (name *of the product*) in (*units: dollars, number of items…*). If you don't understand the product definition, please ask for clarification. You are free to consult any additional (non-historical) information you already have

that might be related to the product and their business development. A system forecast for next month has been produced. This forecast has taken in account three elements: historical trend of data, seasonal effects, and increasing/decreasing effects. Please tell us if you would keep or modify this system forecast for next month, taking into account your judgment and knowledge of the product and the business.

(If the subject wants to modify the forecast) Please indicate how large the modification will be and in what direction. You are free to specify a percentage or a value in units of modification.

(*After the forecast is produced*) Please explain your motivations and reasons for this result.

Appendix B

Credibility of system forecasts scale:

In your opinion, a system forecast is (mark just one option for each question):

Fair/unfair

Biased/unbiased

Tells the whole story/does not tell the whole story

Accurate/inaccurate

Can be trusted/cannot be trusted.

Appendix C

Expertise knowledge subscale

	Completely agree	**Partly agree**	**Neutral**	**Partly disagree**	**Completely disagree**
This person has knowledge that is specific to his or her field of work.					
This person shows that they have the education necessary to be an expert in their field					
This person has knowledge about their field					
This person conducts research related to their field					
This person has the qualifications required to be an expert in their field.					
This person has been trained in his or her area of expertise					