Organizational Behavior and Human Decision Processes 120 (2013) 24–36

Contents lists available at SciVerse ScienceDirect

Organizational Behavior and Human Decision Processes

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

On the relative importance of linear model and human judge(s) in combined forecasting

Matthias Seifert [a,*], Allègre L. Hadida [b,*]

[a] Operations & Technology Area, IE Business School, Maria de Molina, 12, 5th Floor, 28006 Madrid, Spain
[b] Judge Business School and Magdalene College, University of Cambridge, Trumpington Street, Cambridge CB2 1AG, UK

ARTICLE INFO

Article history:
Received 22 March 2010
Accepted 23 August 2012
Available online 4 October 2012
Accepted by William Bottom

Keywords:
Judgmental forecasting
Combined forecasts
Bootstrapping models
Music industry

ABSTRACT

When and to what extent should forecasts rely on linear model or human judgment? The judgmental forecasting literature suggests that aggregating model and judge using a simple 50:50 split tends to outperform the two inputs alone. However, current research disregards the important role that the structure of the task, judges' level of expertise, and the number of individuals providing a forecasting judgment may play. Ninety-two music industry professionals and 88 postgraduate students were recruited in a field experiment to predict chart entry positions of pop music singles in the UK and Germany. The results of a lens model analysis show how task structure and domain-specific expertise moderate the relative importance of model and judge. The study also delineates an upper boundary to which aggregating multiple judgments in model-expert combinations adds predictive accuracy. It is suggested that ignoring the characteristics of task and/or judge may lead to suboptimal forecasting performance.

Introduction

The question of when and to what extent forecasts should rely on analytical models or human judges is fundamental to many organizations and becomes increasingly important as technologies continue to advance. For instance, hedge fund managers have intensified their use of algorithmic trading approaches to determine the timing, price and quantity of trading orders. In some markets, automated trading can lead to faster interpretations of complex business information, enabling analysts to reduce their reaction time to emerging trends. Similarly, IBM announced the first commercial application of their Watson supercomputer to the healthcare industry. Using analytical data mining combined with image and speech recognition software, IBM's technology promises to complement physicians' expertise and help them diagnose and treat cancer patients in a more efficient way. Most recently, 2012 Oscar nominee "Moneyball" opened the "model versus expert" debate to a broader public audience. The movie, which is an adaptation of Michael Lewis's best-selling book, tells the true story of the Oakland Athletics – a baseball team that achieved astounding success on the pitch by relying on a purely mathematical approach to selecting team members (Lewis, 2004).

Yet, failures in the trading algorithms may have disastrous consequences in the financial markets. Cancer patients may be misdiagnosed and may die when relying purely on Watson's analytical output. And although the Oakland As' reached the playoffs in four consecutive years from 2000 to 2003 and won their first playoff in 2006, they have not finished within ten games of the division lead since then – admittedly partly because the team's model-based strategy has now become commonplace in several major league sports, including baseball, where top teams such as the New York Yankees and Boston Red Sox emulated it with much larger budgets. In a newly leveled playing field that is increasingly correcting for the effects of undervalued assets, however, increased stages of sophistication in mathematical modeling may soon yield decreasing returns: "Everyone can look at the same numbers, there are lots of mathematicians for hire, and so secrets are hard to keep" (Cowen & Grier, 2011). These examples show that traders, physicians and sport managers alike need to continuously re-assess when and to what extent they can sensibly base their decisions on model outputs, and when and to what extent they should complement them with their own judgment.

This debate is at the heart of our study. Consistent with prior research in the field of judgmental forecasting, we specifically focus on the value of combining forecasts generated by *linear* models with judgmental predictions. Existing studies have discussed the strengths and weaknesses of using linear models or human judgments on their own (Dawes, Faust, & Meehl, 1989; Lawrence, Goodwin, O'Connor, & Önkal, 2006; Meehl, 1996; Schoemaker, 1993). However, demonstrations of how a combination of both model and judge can lead to higher forecasting accuracy are few and far between (Blattberg & Hoch, 1990; Lawrence, Edmundson,

* Corresponding authors. Fax: +44 (0) 1223 339701 (A.L. Hadida).
E-mail addresses: matthias.seifert@ie.edu (M. Seifert), a.hadida@jbs.cam.ac.uk (A.L. Hadida).

0749-5978/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.obhdp.2012.08.003

& O'Connor, 1986). Among the few existing studies, it has been suggested that a simple 50/50 weighting between model outputs and managerial judgment is likely to outperform either of the two alone (Blattberg & Hoch, 1990). This finding had a profound impact on our understanding of the use of decision support tools and spawned several follow-up studies (for example, Hoch & Schkade, 1996; Sanders & Ritzman, 1995; Stewart, Roebber, & Bosart, 1997). However, none of the subsequent studies to date have attempted to examine the robustness of the proposed 50/50 split in greater depth.

We do so in the present paper. Specifically, we analyze the influence of task structure, domain-specific expertise and aggregated judgments on the effectiveness of combined model-judge(s) forecasts. Our empirical sample draws on real world music industry experts predicting the chart success of upcoming pop music singles. The following section provides an overview of the relevant literature and introduces our main hypotheses. We then present the forecasting task and give details of our empirical setting and data collection methods. The subsequent section provides a comprehensive description of the methods used for conducting the study. The data analysis and results section discuss regression results, with a particular emphasis on their out-of-sample generalizability. In the discussion and conclusions section we summarize key findings, offer alternative explanations of the studied phenomena and outline further research.

Theoretical background and hypotheses development

Model-Individual judge combinations

Mechanical forecasting models, whether they are heuristic or estimated from historical data (Hoch & Schkade, 1996), process information in a consistent, systematic and logical manner (Blattberg & Hoch, 1990). While linear models tend to generate fewer errors than human judges, those that occur are also more likely to be large and to lead to leptokurtic distributions[1] caused, for instance, by the use of inappropriate logical rules (Meehl, 1996; Peters, Hammond, & Summers, 1974). In contrast, judges are proficient in identifying new prediction variables (Blattberg & Hoch, 1990) and in providing subjective assessments of variables that are difficult to measure objectively, such as ethical, moral or aesthetic judgments (Einhorn, 1974). Yet, judges are likely to be biased (Kahneman & Tversky, 1984), influenced by organizational politics (Shapira, 2002), and inconsistent when using information.

Judgmental forecasting research has therefore frequently led to the conclusion that predictions based on model-expert combinations prove to be superior to model outputs and expert judgment alone (e.g., McClish & Powell, 1989; Sanders & Ritzman, 1995; Stewart, 1997; Yaniv & Hogarth, 1993). Individual buyers' predictions of catalog sales and individual brand managers' forecasts of coupon redemption rates provide two specific contexts in which the predictive value of model-judge combinations was investigated (Blattberg & Hoch, 1990). In both settings, the authors not only demonstrate how the aforementioned 50:50 split between model and judge leads to higher forecasting accuracy. They also offer a more refined understanding of the way in which the two forecasting inputs interact. In particular, the complementarity of model and judge seems to be directly linked to judges' ability to process task information in a nonlinear way. Taking this into account, the present study sets out to test a variety of task- and judge-related factors that may influence the robustness of Blattberg and Hoch's (1990) initial findings.

[1] A leptokurtic or "super Gaussian" distribution is defined as a statistical distribution with a positive excess kurtosis.

Task structure

Proponents of Brunswik's (1956) probabilistic functionalism share the view that we live in an objective world in which uncertainty resides only in the minds of decision makers. The degree to which a task is perceived as well- or ill-structured is therefore subjective, and relates to the cognitive abilities of the decision maker.

In this paper, we define "task structure" relative to the level of validity and reliability of the cue contents of a specific task. Ill-structured tasks arise through environmental changes that impact the probabilistic linkages between task input and outcome (Wood, 1986). Our conceptualization shares the commonly adopted view that environmental changes frequently lead to missing information relevant to the task (Fellner, 1961; Frisch & Baron, 1988). In turn, missing information reduces the transparency of means-ends relationships.

This effect is evident in unknown probability distributions linking informational cues and task outcome (Camerer & Weber, 1992), in unknown path sequences, and in the absence of appropriate algorithms for integrating task information (Deng, 1996; Hammond, 1996; Simon, 1977; Steinmann, 1976; Wood, 1986). For instance, when forecasting in disruptive industry environments, a sudden regime change can lead to obsolete historical sales data. Uncertainty may consequently arise regarding the relationship between informational cues and the forecasting event, and, hence, regarding the general type of model that is most appropriate to generate accurate forecasts.

Domain-specific expertise

Expert knowledge is highly organized, domain-specific and not transferable (Feltovich, Prietula, & Ericsson, 2006; Hogarth, 2001). It allows judges to process information more quickly and achieve a higher degree of accuracy than novices (Blattberg & Hoch, 1990; Klein, 2003).

Experts' proficiency in utilizing contextual information is particularly salient when generating forecasts in the context of ill-structured tasks (Armstrong, 1983; Lawrence, Goodwin, O'Connor, & Önkal, 2006; Spence & Brucks, 1997). In fact, research in judgmental forecasting showed that experts were likely to outperform linear models when forecasting ill-structured tasks, thanks to their superior ability to anticipate sudden changes in the data structure (Sanders & Ritzman, 1995). These sudden changes may result from nonlinear relationships among informational cues (Kleinmuntz, 1990; Sanders & Ritzman, 1995; Yaniv & Hogarth, 1993). Hence, experts are likely to select fewer, more diagnostic cues from the contextual knowledge surrounding the task and evaluate these cues more consistently than novices and models (Alexander, 1995; Armstrong, 1983; Spence & Brucks, 1997). Data structures of well-structured tasks are less variable and based on linear relationships between informational cues. The latter enable experts to generate accurate forecasts without having to evaluate contextual information.

We consequently anticipate a positive relationship between the amount of contextual knowledge needed to achieve high forecasting accuracy and the degree to which the forecasting task is ill-structured. Based on the proficiency of experts to utilize contextual knowledge, we also conjecture that the optimal model-judge combination shifts towards a heavier reliance on human judgment when tasks are ill-structured. In contrast, well-structured tasks can be described fairly accurately in terms of linear relationships between informational cues and require proportionally less contextual knowledge to generate forecasts. Optimal model-judge combinations will therefore primarily rely on model components in well-structured task contexts, regardless of domain-specific expertise. In sum, Hypothesis 1 proposes a moderation effect of task structure and domain-specific expertise on the relative importance of linear model outputs and individual human judgments in

forecasting accuracy. For the sake of clarity, the single Hypothesis 1 is broken down below into three sub-hypotheses (Hypotheses 1a, 1b and 1c).

Hypothesis 1a. When the task is well-structured, combined model-judge forecasts will rely more on linear model outputs than on human judgment, regardless of the level of domain-specific expertise.

Hypothesis 1b. When the task is ill-structured and domain-specific expertise is low, combined model-judge forecasts will rely more on linear model outputs than on human judgment.

Hypothesis 1c. When the task is ill-structured and domain-specific expertise is high, combined model-judge forecasts will rely more on human judgment than on linear model outputs.

Model-multiple judge combinations

Aggregated group judgments draw on a greater variety of knowledge sources. As such, they are likely to outperform individual judgments, because individuals included in the group may explain different parts of the task variance (Lawrence, Goodwin, O'Connor, & Önkal, 2006). In line with the "wisdom of crowds" effect (Surowiecki, 2004) according to which combined predictions of multiple individuals are likely to eliminate judgmental inconsistencies and biases (Einhorn & Hogarth, 1975; Einhorn, Hogarth, & Klempner, 1977; Goldberg, 1970), it is desirable to include individuals with diverse levels of expertise in the group (Bunn, 1985; Goodwin, 2000). To date, Stewart et al. (1997) provided the only empirical insights into the effectiveness of model-multiple judge combinations. By computing the mean of four random judgments, they demonstrated how an aggregation of judges was likely to pick up a larger portion of unmodeled variance, resulting in higher overall accuracy of the model-multiple judge combination.

Two important questions however remain unanswered. First, as the study was conducted in a relatively stable environment where model forecasts explained more than 95% of the variance, we still do not know much about how accurate group judgments prove to be when dealing with ill-structured forecasting tasks. Second, it is unclear how the aggregation method itself would influence the effectiveness of model-multiple judge combinations.

Common sense would dictate that the aggregated group judgment of two or more experts is likely to exploit contextual knowledge more effectively and with higher forecasting accuracy than one individual expert judgment alone. Based on our previous discussion of the influence of task structure on model-individual judge combinations, we argue that task structure will similarly moderate model-multiple judge combinations. In particular, the predictive power of mechanically aggregated group judgments combined with model outputs is likely to account for a larger portion of unexplained model variance than individual model-judge combinations in highly complex and ambiguous (ill-structured) task environments. Conversely, when task structures are associated with relatively clear and stable causal relationships, models may be fairly accurate representations of the task environment, and aggregated judgments will have a lesser positive influence on predictive accuracy.

Hypothesis 2. Model-multiple judge combinations outperform model-individual judge combinations to a greater extent when the task is ill-structured than when the task is well-structured.

Consider next the optimal number of experts necessary to enhance the performance of combined model-multiple judge forecasts. The finding that the predictive accuracy of mechanically aggregated judgment follows a concave function, with incremental improvements diminishing after twelve aggregated forecasts and the largest marginal improvement occurring during the aggregation of the first few judgments (Ashton, 1986; Hogarth, 1978), appears fairly robust. It holds true in different task environments and even when including the most inaccurate experts in the sample. The robustness of this phenomenon when combining multiple human judges with linear models has not yet been subject to empirical investigation.

In particular, a systematic analysis of how group size influences the accuracy of model-multiple judge combinations may shed light on whether an increase in group size can lead to a more efficient interpretation of the nonlinear relationships among informational cues in both ill- and well-structured task contexts. Although contextual knowledge, as stated in the first two hypotheses, may be of greater value in ill-structured environments, judges are likely to utilize contextual information in well-structured settings as well, albeit to a lesser extent. The relative importance of human judgment may therefore increase consistently with the number of judges included in model-multiple judge combinations. Even so, an aggregate of individual expert judgments is unlikely to account for all of the nonlinear relationships among informational cues. We therefore hypothesize a tipping point at which the addition of one more judgment to the group aggregate ceases to improve predictive accuracy.

Hypothesis 3. When combining model and multiple judges, an increase in the number of aggregated judgments will yield a marginally diminishing improvement in forecasting accuracy in both ill- and well-structured task environments.

Empirical context and data collection

Empirical context

The music industry forms our empirical setting. It is a prime example of a disruptive environment, where enabling technologies and consumer preferences change rapidly and continually. We focus on the two largest domestic markets within the European Union: Germany and the United Kingdom. These two countries have no regulatory quota systems and have therefore comparable domestic markets.[2] Berlin and London are also the locations of the global headquarters of Universal Music and EMI, two of the "big four" global record companies (IFPI, 2011).[3]

The music industry provides an appealing context to study model-judge combinations, especially at a time when the pendulum seems to be swinging toward a prevalence of analytical decision support models: "The power and the decision has sat with the A&R man,[4] who gets up late in the day, listens to lots of music, goes to clubs, spends his time with artists and has a knack of knowing what would sell. (...) What we are doing is taking the power away from the A&R guys and putting it with the suits- the guys who systematically work out how to sell music" (Guy Hands, former CEO of EMI Music Group, quoted in Gapper, 2008).

The music industry setting allows us to measure predictions based on real-world time series data rather than laboratory results.

[2] Such quotas oblige local companies to produce, distribute and/or broadcast a certain percentage of domestic music products. They may be seen as attempts to preserve cultural heritage, and are a common regulatory response to perceived cultural threats from abroad. Their restrictions typically affect national radio and television broadcasting. For instance, 40% of programs broadcast through French media channels must be of French origins, and similar quotas exist in Spain (Cohen, 1993).

[3] The "big four" (Warner Music Group, Sony BMG Music Entertainment, Universal Music and EMI) jointly account for 72% of the global music market. Among them, Universal Music and EMI account for 37% of the global music market (IFPI, 2011).

[4] Artist & Repertoire manager, responsible for scouting new talents.

Our forecasting event under investigation is the entry position of pop singles in the national Top 100 charts.[5] Alongside traditional sales target estimates, A&R managers routinely predict chart entry positions to assess the success potential of a single. These predictions are difficult to make, as a successful single must satisfy the needs of two target groups: the media (to increase visibility of the single and get it promoted) and end-customers (to get them to purchase and listen to the single).

The actual chart entry of a pop single is usually preceded by a promotion period of 8–10 weeks. During this promotion period, the single may be broadcast on radio and television and may appear in print and online media or as part of a promotional retail campaign. As record companies must register all planned release dates with a central authority at least 3 months in advance, release schedules represent public information that record companies use to create release strategies aimed at optimizing sales –for instance, by avoiding simultaneous releases of similar singles from their own or rival labels. Singles featured among the 100 bestselling songs in one given week are included in the charts with a 1-week lag.[6] Singles tend to reach their peak position on the day of their chart entry, which reflects accumulated sales during their first week. This observation is consistent with research findings in other cultural industries such as cinema (see for instance Zufryden, 1996).

Due to the blockbuster nature of the music industry, record companies typically hold large portfolios of artists, with a minority of them generating profit (Vogel, 2004). Forecasting the success of new artists entails a substantial amount of uncertainty about market performance and may therefore be described as an ill-structured task. Conversely, the success of artists with an established, historical track record may be easier to anticipate, and its prediction may be described as a well-structured task.

Data collection

In order to identify the variables to be included in our models and to ensure that said models would accurately represent experts' predictions of chart entry positions in the music industry, we conducted 23 in-depth semi-structured interviews in Germany and in the UK with senior record companies' A&R managers. We selected respondents based on their level of industry experience and on their hierarchical position in their organization, and contacted them through the leading international professional association 'HitQuarters'. All 23 respondents were managers with divisional or regional responsibilities who had at least 10 years of industry tenure.

The interviews led to the identification of three categories of variables relevant to chart predictions. The first relates to the promotion of the single (airplay chart position, marketing expenses, price, release date, and media attention) and corroborates findings that promotional activities have a crucial impact on the performance of a music product (Moe & Fader, 2001). The second indicates the artist's previous success record (number of major awards won and average chart position of previous singles).[7] The third reflects contextual knowledge of the single itself (song title, artist's name, producer's name, songwriter's name, artist's picture, and audio sample).

We used these predictor variables to design an online questionnaire displaying a set of as-yet unreleased pop singles, and asked respondents to estimate the chart entry position of these singles. The online questionnaire described each single through a cue profile that included the three categories of variables relevant to chart predictions detailed above. That is, data related to the promotion of the single, including a photograph of the artist; data related to the artist's previous success record; and nonlinear, contextual knowledge of the single itself, including a 30-s sound sample that could be directly listened to by clicking on a provided link.

We drew on information obtained from marketing research firms, chart compilation companies, retailers, key media firms, record companies and the internet to configure cue profiles. The questionnaire comprised pop singles from both newcomers and established artists. These two groups differ in the number of predictor variables available to make a forecasting judgment. In particular, the number of previously released albums and singles and the number and nature of official sales level certificates awarded in recognition of the quantity of albums or singles sold distinguish newcomers from established artists.

We asked 180 respondents to make quantitative predictions on the singles' chart entry position. Eighty-eight postgraduate students at two major universities in Germany and in the UK comprised our "Novices" control sample. We incentivized them by entering them into a drawing for book vouchers. We also recruited industry professionals by offering feedback on their own performance when predicting success and a full industry-specific report summarizing our results. Interviews revealed that a fairly robust measure of the level of expertise in the music sector relates to whether music managers work in major or independent record labels. Of the 92 A&R managers who participated in the study, 44 were based in the UK, and 48 in Germany. Forty-eight worked for the "big four" major record companies and 44 for 17 small and medium-sized independent labels in both countries.

To increase sample representativeness, we only approached the 17 largest and most successful independent labels as identified by domestic associations for independent music (HitQuarters, 2008). Ten of these independent labels were located in Germany, and seven were located in London. A large number of independent labels are not publicly traded, making their global market shares difficult to assess. To ensure that respondents had some level of domain-specific expertise, we measured success in this context as the frequency of placing artists in the Top 100 charts, which we defined as their "hit rate." Managers in independent labels are often specialized in particular music genres, have access to much smaller budgets, and are typically at the beginning of their career. Many of them also aspire to joining a major record label as they gain industry experience and create a track record of chart successes. There are strong incentives for managers from independent labels to join major record companies and for major labels to poach the most successful managers from independent labels.

We therefore classified ex ante our 180 respondents according to their familiarity with the task environment (Cooksey, 1996), which we captured as the extent of their industry expertise. Managers from major record companies (hereafter referred to as "Major Label Experts") were highly familiar with the task environment. They displayed a broad domain-specific expertise and a high degree of task-related experience, manifested in long firm tenure and a relatively high hit rate of single releases. Managers from independent labels ("Independent Label Experts") had narrower domain-specific expertise, which was often limited to particular music genres. They tended to be less experienced, and less successful in producing "chart breakers." Our control group of postgraduate students (Novices) had the lowest degree of experience and familiarity with forecasting hit singles.

Data collection took place in four stages over 12 weeks (February to May 2007). We designed our questionnaires to allow periods of 2–3 weeks between prediction and actual chart entry,

[5] A single is defined as a song extracted from a current or upcoming album and released on its own, often to generate subsequent sales of said album.

[6] The earliest singles can appear in the charts is therefore the week after they have been made available for retail.

[7] Twenty of our interviewees noted that artists' chart histories, inasmuch as they illustrate the size of their fan base and their commercial experience, provide a fairly reliable and sustainable indicator of their current and future success.

equivalent to the periods of 1–2 weeks between prediction and release date. This ensured that every single was advertised for the same duration. The fact that, on average, only five to ten singles enter the Top 100 charts every week justifies a four-stage data collection.

We generated 105 cue profiles for each country, leading to a total of 210 prediction cases altogether. Every respondent, expert and novice alike, generated forecasts for 40 pop singles. In addition to raw predictions of success, Novices were also asked to provide information about their music awareness and consumption on five-point Likert scales, from 1 (lowest) to 5 (highest level of music trend awareness). We randomly assigned pop singles from newcomers and established artists to respondents. Participants were given 1 week to return the questionnaires, and the response rate was 72%.

Methods

Our general research design is based on a Brunswikian lens model used in judgment analysis (Brunswik, 1956). We also test whether task structure and domain-specific expertise moderate the optimal relative importance of linear model outputs and individual judgment (Hypotheses 1a–c) by pooling Major Label Experts and Independent Label Experts and performing a 2×2 factorial ANOVA. We rely on "task structure" and "domain-specific expertise" as the independent variables and the "optimal relative split between model and judge" as the dependent variable.

Hypothesis 2 proposes that model-multiple judge combinations outperform model-individual judge combinations to a greater extent when the task is ill-structured than when the task is well-structured, and Hypothesis 3 posits a marginally diminishing improvement in predictive accuracy as the number of human judges involved in model-multiple judge combinations increases. The tests of Hypothesis 2 and Hypothesis 3 are conducted by relying on hierarchical regression models.

Lens model analysis

We followed a regression procedure first introduced by Blattberg and Hoch (1990) to assess the predictive accuracy of a linear mathematical model combined with one human judgment. We extend the original model by incorporating mechanically aggregated group judgment in the forecasting combination. Our key interest lies in understanding judges' ability to extract predictive value from contextual information associated with the forecasting problem, thereby improving the accuracy of linear models considered on their own (see also Blattberg & Hoch, 1990; Hoch, 1987, 1988; Stewart et al., 1997).

The lens model consists of a specific target event (Y_e), which in our context is defined as the chart entry position of a single, resulting from the configuration of multiple and potentially redundant observable linear environmental cues (Xi), which in our context are defined as the predictor variables included in the online questionnaire (Cooksey, 1996; Hammond, 1996). The best fitting model of target event (Y_e) accounts for the weight (β_e's) associated with each cue (X_i) in the task environment and for the standard error of the model (ε_e), as follows:

$$Y_e = X_i\beta_e + \varepsilon_e. \tag{1}$$

Respondents observe linear and nonlinear cues before making a forecasting judgment (Y_s) about the chart entry position of a single (Y_e). The best fitting model of the environment delineates an upper boundary of the predictive accuracy that forecasters are able to achieve from interpreting cues (X_i) in a linear way. If the regression function of target event (Y_e) is used to generate predictions about a new set of estimation data ($\hat{Y}_e$), then the model becomes:

$$\widehat{Y}_e = X_i\hat{\beta}_e, \tag{2}$$

where $\hat{Y}_e$ represents the best fitting linear model of the environment.[8] The residual portion of the forecaster's judgment beyond the predictions generated by the linear model of $\hat{Y}_e$ may be associated with the domain-specific expertise that enables judges to pick up cues not included in the model or to integrate information in a nonlinear manner. The equation isolates this residual predictive value by regressing the respondent's forecasting judgment (Y_s) onto model predictions ($\hat{Y}_e$), as follows:

$$Y_s = \gamma\widehat{Y}_e + U, \tag{3}$$

where the residual (U) contains both valid expertise and random error. The full model of the respondent's forecasting accuracy with all variables in standardized form then becomes (as per Hoch, 1987 and Blattberg & Hoch, 1990):

$$r_a(y_e, y_s) = \alpha r(y_e, \hat{y}_e) + \sqrt{1-\alpha^2}\, r(y_e, u), \tag{4}$$

where α represents the correlation between y_s and $\hat{y}_e$, $r(y_e, \hat{y}_e)$ describes the accuracy of the linear model, and $r(y_e, u)$ represents the validity of individual judgment.

Combining linear model predictions and individual judgment

Regression function (4) forms the basis for combining model predictions ($\hat{y}_e$) and judgmental forecasts (y_s). The purpose of this combination is to test whether model predictions and human judgments together can achieve a higher degree of accuracy than each of them on their own when facing a new set of estimation data (Blattberg & Hoch, 1990). The optimal combination ($\hat{y}$) of model predictions ($\hat{y}_e$) and judgmental forecasts (y_s) is formulated as:

$$\hat{y} = b_1\hat{y}_e + b_2 y_s, \tag{5}$$

with an overall model fit of:

$$R^2(y_e, \hat{y}) = b_1 r(y_e, \hat{y}_e) + b_2 r(y_e, y_s), \tag{6}$$

where b_1 and b_2 represent relative weights for linear model and individual judge. When integrated with (5), the relative weighting formula becomes:

$$R^2(y_e, \hat{y}) = (b_1 + b_2\alpha) r(y_e, \hat{y}_e) + b_2\sqrt{1-\alpha^2}\, r(y_e, u), \tag{7}$$

where $R^2(y_e, \hat{y})$ represents the optimal fit of the model-judge combination. We use the following hierarchical regression function to isolate nonlinear expertise in model-judge combinations (Blattberg & Hoch, 1990):

$$R^2(y_e, \hat{y}) = R^2(y_e, \hat{y}_e) + R^2(y_e, u), \tag{8}$$

where $\hat{y}$ represents the most accurate possible aggregation of model prediction and individual judgment, and y_e indicates the actual entry chart position of the single. While $\hat{y}_e$ describes the artificially-generated prediction of the best fitting environmental model, the residual u contains both nonlinear expertise and the standard error component of the model. Whenever $r(y_e, u \neq 0)$, then $r(y_e, \hat{y}) \geqslant r(y_e, y_s)$ and $r(y_e, \hat{y}) \geqslant r(y_e, \hat{y}_e)$, which indicates that the combination of model and individual judge is more accurate than model or judge alone. The hierarchical regression function also allows a closer examination of regular partial correlations, where (y_e) and (y_s) are both controlled for linear model predictions ($\hat{y}_e$) and where

[8] In the remaining sections of this paper we use "^" to denote any kind of artificially constructed (rather than naturally observed) forecast, such as linear model outputs or aggregated predictions.

semi-partial correlations only account for expert judgment (y_s). While semi-partial correlations confirm the validity of forecasters' expertise (Cohen & Cohen, 1975), squared regular partial correlations describe the amount of unexplained model variance these forecasters picked up (Hoch, 1987).

Combining linear model predictions and aggregated group judgment

We investigate the value of aggregated group expertise in a model that combines linear model predictions ($\hat{y}_e$) with multiple forecasting judgments (y_N). The first step involves describing a function ($\hat{y}_g$) that aggregates the individual judgments of *N* forecasters. Based on the literature on combining forecasts, we tested the value of forecasting expertise by using four different types of well-established aggregation mechanisms (see for instance Clemen, 1989; de Menezes, Bunn, & Taylor, 2000): (1) simple average, (2) outperformance, (3) optimal, and (4) regression weights. All these methods combine judgmental forecasts via a linear weighting function.

Simple average

Following an equal weighting procedure (Einhorn et al., 1977; Goldberg, 1970; Silverstein, 1987), the simple average aggregation method formulates mechanically aggregated group judgment $\hat{y}_g$ as:

$$\hat{y}_g = \frac{1}{N} y_{s1} + \cdots \frac{1}{N} y_{sN}. \tag{9}$$

Outperformance

The second aggregation technique assigns Bayesian probabilities to each forecast, representing the likelihood that it will perform best in the group (Bunn, 1975). Probabilities serve as individual weights in the aggregation process and are revised after each judgment. They can be computed in terms of the smallest absolute error that a forecaster generated in the past, and they are particularly suitable when relatively little historical data is available or when expert judgment needs to be integrated into the aggregation process (Bunn, 1987; de Menezes et al., 2000). For two judgmental forecasts, the combined prediction can be written as:

$$\hat{y}_g = k_i y_{s1,i} + (1 - k_i) y_{s2,i}, \tag{10}$$

where k_i represents the subjective probability that a forecast will be superior when making judgment *i*. In particular, *k* can be interpreted as a fraction of the total number of previous forecasts in which a judge has been associated with the smallest absolute error– that is, when a judge has been the most accurate (De Menezes et al., 2000). The posterior value of *k* is recalculated after each forecast.

Optimal

The third aggregation method computes individual weights for each forecast by relying on least squares regression, where the constant is suppressed to zero and beta weights equal 1 (Bates & Granger, 1969). The objective of the optimal aggregation technique is to minimize error variance of the forecast combination.

Regression weights

Our final approach to combining judgments relates to ordinary least squares regression, with unstandardized regression weights assigned to each individual forecaster and the constant included when generating combined forecasts. When replacing y_s with $\hat{y}_g$ in Eq. (4), the combined predictive accuracy becomes:

$$r_a(y_e, \hat{y}_g) = \alpha_i r(y_e, \hat{y}_e) + \sqrt{1 - \alpha_i^2}\, r(y_e, \tilde{u}), \tag{11}$$

where $r(y_r, \tilde{u})$ represents the correlation between the residuals of the aggregated group judgment and target event (y_e). Eqs. (5) and (6) subsequently become:

$$\tilde{y} = b_1 \hat{y}_e + b_2 y_g, \tag{12}$$

and

$$R_2(y_e, \tilde{y}) = b_1 r(y_e, \hat{y}_e) + b_2 r(y_e, \hat{y}_g), \tag{13}$$

where $\tilde{y}$ represents a hierarchical regression function in which the predictions of the linear model are entered first and followed by an aggregated group judgment $\hat{y}_g$. Likewise, $R^2(y_e, \tilde{y})$ indicates the model fit when combining linear model predictions with aggregated group judgment. The fully combined model results from inserting Eq. (11) into Eq. (13) as follows:

$$R^2(y, \tilde{y}) = b_1 r(y_e, \hat{y}_e) + b_2(\alpha_i r(y_e, \hat{y}_e) + \sqrt{1 - \alpha_i^2}\, r(y_e, \tilde{u}). \tag{14}$$

The orthogonality of $\hat{y}_e$ and u allows for a unique variance partitioning between linear model and human judge. It allows the simplification of function (14) as follows:

$$(y_e, \tilde{y}) = R^2(y_e, \hat{y}_e) + R^2(y_e, \tilde{u}), \tag{15}$$

We posit that $\tilde{u}$ may be used to investigate the value of aggregated group judgment.

Task structure: distinguishing between ill-structured and well-structured tasks

Task structures have typically been operationalized as artificially constructed scores using an equal unit weighting procedure to account for the key properties of a given task environment (Dunwoody, Haarbauer, Mahan, Marino, & Tang, 2000; Hammond, 1988; Stewart et al., 1997). We measure task structure relative to two key factors: cue validity and cue reliability. Cue validity relates to the environmental predictability when using relevant cues and can be measured by the R^2 value of the best fitting model in the environment (Steinmann, 1976). Cue reliability is measured in terms of the mean inter-correlation of cue judgments provided by several experts. The resulting task structure score (TSS) can be formalized as:

$$TSS = (10 * (1 - R^2) + 10 * (1 - r))/2, \tag{16}$$

where the first part of the numerator ($10 * (1 - R^2)$) implies that low cue validity indicates an ill-structured task and high cue validity indicates a well-structured task. The second part ($10 * (1 - r)$) refers to our measure of cue reliability, and low inter-correlation again implies high ill-structuredness. In line with Dunwoody et al. (2000), the multiplier 10 is simply used as a scaling variant for the final task structure score.

In sum, the first step in our research design entails the use of a hierarchical regression model, which allows us to closely examine the predictive performance of both linear model and judge(s). In the second step, we use the outcome of the hierarchical models as our dependent measure and compute task structure scores that serve, together with participants' level of domain expertise, as our independent measures. The resulting variables subsequently represent the basis for a 2×2 factorial ANOVA.

Analysis and results

Our analysis and results section is organized as follows. We first report on baseline analyses of the level of success when manipulating task contexts, the predictive accuracy of models and judges alone, and the similarity of our model-judge combinations to other existing studies (Blattberg & Hoch, 1990; Stewart et al., 1997).

These baseline analyses are prerequisites to the test of our hypotheses, and help position our approach and findings relative to other empirical research on judgmental forecasts. The subsequent test of our hypotheses focuses in particular on the interactions between task structure and domain-specific expertise.

Baseline analyses

Task structure score

Our field data confirm that pop music singles of newcomers and established artists relate to two distinct task structures. First, the best fitting model of the task environment when forecasting the success of newcomers is associated with a smaller R^2 value, indicating lower environmental predictability in comparison to well-structured tasks (ill-structured task: $R^2_{ill} = 0.27$; well-structured task: $R^2_{well} = 0.51$). Second, the two tasks also differ with regards to cue reliability. Specifically, our interview data show that experts agree to a greater extent on the appropriateness of cues when predicting the success of established artists than when predicting the success of newcomers. This is indicated by lower mean cue intercorrelations in the case of ill-structured tasks than in the case of well-structured tasks ($r_{ill} = 0.36$; $r_{well} = 0.61$, $p < .01$). The differences in environmental predictability and cue reliability translate into two distinct Task Structure Scores (ill-structured task: $TSS_{ill} = 4.40$; well-structured task: $TSS_{well} = 6.85$).[9]

Performance of model and judge alone

On average, Novices were 24 years old. They estimated their mean annual expenditures on music-related products at 84.8 GBP, and their awareness of current music trends at a mean value of $M = 2.45$ ($SD = .57$) on a 5-point Likert scale ranging from 1 (lowest) to 5 (highest level of music trend awareness). As anticipated, the average industry tenure and success rate of Major Label Experts and Independent Label Experts differed quite substantially. The former had a mean industry tenure of 11.1 years ($SD = 3.6$) and placed, on average, 15.3 singles in the Top 100 charts. The latter had a mean industry tenure of 7.6 years ($SD = 5.6$) and placed, on average, 2.6 singles in the Top 100 charts.[10]

To improve out-of-sample forecasting predictability, we controlled for model shrinkage (Cooksey, 1996) of both model and judge components. In particular, we performed cross-validation analyses by using half of our dataset to draw ten random samples and simulated predictions about the remaining data. Results indicate a relatively good fit between experts' forecasts of the chart entry position of a single and its actual chart entry position (mean $R^2 = 0.46$), despite an average shrinkage of 14%. Novices' judgments were less precise. The fit between their forecasts and actual entry position was reduced to $R^2 = .23$, and led to a model shrinkage of approximately 20%. Table 1 below offers an overview of cross-validated model fits.

Table 2 shows that when comparing the performance of a linear model to that of a human judge (regardless of the forecaster's expertise), the model achieved a higher predictive accuracy in well-structured task conditions (that is, when forecasting the chart entry positions of established artists: model accuracy: $r = .69$; judge accuracy: $r = .59$). However, in ill-structured task conditions (that is, when forecasting the chart entry positions of newcomers), an expert judge achieved a higher predictive accuracy than the linear model (model accuracy: $r = .44$; expert judge accuracy: $r = .52$). Novices displayed the lowest predictive accuracy in both task environments (ill-structured task: $r_{ill} = .36$; well-structured task: $r_{well} = .39$).

Performance of model and judge combined

We regressed real chart positions (Y_e) onto statistical model predictions ($\hat{Y}_e$) and individual respondents' forecasts (Y_s). In line with Blattberg and Hoch's (1990) findings and as shown in Table 2, combining model and judge tends to result in both higher predictive accuracy and improved model fit than when considering model or judge alone. This is highlighted by the mean difference (Δ) between model-judge combination and the better of the two individual decision inputs (R^2 linear model/R^2 human judge).

We then analyzed the predictive power of human judgment by computing semi-partial correlations between real chart positions (Y_e) and residuals (u) and regular partial correlations between chart positions and respondents' predictions, controlling for the linear model. When using semi-partial correlations, all respondents seemed to be able to pick up at least some of the nonlinearities in the task ecology (mean: $r(y_e, u) = .31$).

Experts are associated with a higher average semi-partial correlation ($r(y_e, u) = .38$) than Novices ($r(y_e, u) = .23$). In terms of squared regular partial correlations ($r(y_e, y_s \cdot \hat{y}_e)^2$), Major Label Experts account for 9.5% of the variance unexplained by the linear model across task structures. In contrast, Independent Label Experts, on average, account for only 5% of this variance, and Novices, our control group, are even less effective ($\Delta = .02$).

Table 3 also shows that when converting standardized beta coefficients of each model-judge combination into relative weights (w), our results only reinforce the proposed 50/50 model-expert split when Major Label Experts forecast the chart entry position of established artists (model: $w = .49$; Major Label Expert: $w = .51$). In contrast, when forecasting the chart entry position of newcomers, the emphasis shifts towards the individual judgment of Major Label Experts (model: $w = .28$; Major Label Expert: $w = .72$). For Independent Label Experts, the optimal combination changes from an emphasis on expert judgment when forecasting the chart entry position of newcomers (model: $w = .44$; Independent Label Expert: $w = .56$) to an emphasis on model predictions when forecasting the chart entry position of established artists (model: $w = .65$; Independent Label Expert: $w = .35$).

Hypothesis testing

Hypotheses 1a–c posit a moderation effect of task structure and domain-specific expertise on the relative importance of linear model outputs and individual human judgments in forecasting accuracy. We find that this is indeed the case. The results of the univariate analyses carried out to test Hypotheses 1a–c indicate two main effects and a significant interaction between the predictor variables.

In particular, and as predicted by Hypothesis 1c, an increase in the level of domain expertise led to a significant increase in the relative importance of judge versus model (main effect 1). In fact, the associated level of expertise determined whether or not the optimal model-individual judge combination should primarily rely on model or judge (optimal model-novice combination: 61/39; optimal model-expert combination: 46/54; $F(1,359) = 28.780$, $p < .01$, partial $\eta^2 = 0.159$).

Similarly, the second main effect associated with task structure suggests that the degree of ill-structuredness had a decisive impact on whether the optimal combined forecast should mainly rely on model or judge. Specifically, in well-structured tasks, the optimal split between model and judge (regardless of level of expertise)

[9] We also tested alternative models of the task environment by relying on the set of variables interviewees deemed relevant to predict chart entry positions. In particular, we used stepwise selection to investigate the model fit of all possible combinations of variables to identify the best linear models for the two task environments under consideration (Mentzer & Moon, 2005). The resulting models were based on seven predictor variables for ill-structured tasks ($R^2 = 0.29$) and ten predictor variables for well-structured tasks ($R^2 = 0.50$).

[10] We did not find significant differences between forecasts from respondents based in the UK and in Germany.

Table 1
Cross-validated results: model, judge, and model-individual judge combination.

Task structure	Type of judge	R^2 model	R^2 judge	R^2 model + judge combination	Δ	Validity of expertise $r(y_e, u)$	Variance picked up by judge $r(y_e, y_s \cdot \hat{y}_e)$ (%)
Ill-structured task	Expert (major)	.27	.53	.61	.08	.56	16
	Expert (indep)	.27	.39	.41	.02	.40	7
	Novice	.27	.26	.29	.02	.25	2
Well-structured task	Expert (major)	.51	.51	.64	.13	.36	3
	Expert (indep)	.51	.48	.52	.01	.20	3
	Novice	.51	.18	.52	.01	.21	2

Table 2
Correlation Table showing Model and Judge Achievements.

Type	Ill-structured task		Well-structured task	
	Chart position (y_e)	Model forecast ($\hat{y}_e$)	Chart position model (y_e)	Model forecast ($\hat{y}_e$)
Model (m)	.44	1.0	.69	1.0
Experts (major label)	.57	.29	.66	.60
Experts (indep. label)	.45	.31	.54	.57
Novices	.36	.30	.39	.30

Notes: Table shows correlations between the real chart position of a single (y_e), model prediction ($\hat{y}_e$) and human judgment. All correlation coefficients represent mean values of each respondent group. Major) Label Experts: N = 44; Independent Label Experts: N = 48; Novices: N = 88. Significance levels for all results: $p < .01$.

was 58/42. Ill-structured tasks generally lead to heavier reliance on the judgmental component (optimal model-judge split: 43/57; $F(1,359) = 14.378$, $p < .01$, partial $\eta^2 = 0.086$). Our results also showed a significant interaction effect between domain expertise and task structure ($F(1,359) = 13.310$, $p < .01$, partial $\eta^2 = 0.081$). In particular, the resulting interaction graph (Fig. 1) illustrates that combined model-judge forecasts in well-structured environments are based primarily on model output, regardless of the level of domain expertise (relative weight of human judgment: novice: $w = 0.38$, $SD = 0.20$; expert: $w = 0.43$, $SD = 0.14$). This finding shows consistency with Hypothesis 1a, according to which combined model-judgment forecasts will rely more on a linear model than on human judgment when the task is well-structured, regardless of the level of domain-specific expertise.

The full moderation effect can be revealed when considering forecasts in ill-structured environments: While a combination of low expertise and ill-structuredness leads to heavier reliance on model than on judge (relative weight of human judgment: $w = 0.38$, $SD = 0.18$), high expertise shifts the emphasis towards primary reliance on a judge (relative weight: $w = 0.65$, $SD = 0.17$). Our findings, therefore, support Hypothesis 1b, which states that when the task is ill-structured and domain-specific expertise is low, combined model-judgment forecasts will rely more on linear models than on human judgment. They also support Hypothesis 1c, according to which when the task is ill-structured and domain-specific expertise is high, combined model-judgment forecasts will rely more on human judgment than on linear models.

Aggregated judgments were built separately for Major Label Experts and Independent Label Experts and involved two to twelve managers. For each number of forecasters associated with a particular aggregated judgment, we used a simple unit weighting technique to combine individual predictions (Cooksey, 1996). Hypothesis 2 postulates that the benefit of relying on aggregated rather than individual judgment is stronger when facing ill-structured forecasting tasks. This is effectively the case. We tested Hypothesis 2 by analyzing how an increase in group size impacts (a) the validity of expertise and (b) the percentage of unmodeled variance picked up by forecasters. Consistent with previous research on opinion aggregation (Ashton, 1986; Hogarth, 1978), predictive accuracy, validity of expertise and unexplained variance picked up by subjects are associated with concave functions (Fig. 2).

Table 3
Relative Importance of Model versus Judge Components.

	Well-structured		Ill-structured	
	Model	Judge	Model	Judge
Experts (major label)	0.49	0.51	0.28	0.72
Experts (indep. label)	0.44	0.56	0.35	0.65

Major Label Experts and Independent Label Experts marginally add judgment validity and pick up a greater proportion of unexplained model variance when the task is ill-structured. Major Label Experts showed an incremental increase in validity of expertise of $r(y_e, u) = .07$ when forecasting the chart entry position of newcomers and $r(y_e, u) = .05$ when forecasting the chart entry position of established artists. Furthermore, every additional Major Label Expert included in the model-multiple judge combination picked up about 11.3% of unmodeled variance when considering established artists and 5% when considering newcomers. On average, Independent Label Experts picked up a slightly lower percentage of unexplained variance (ill-structured tasks: $(r(y_e, y_s \cdot \hat{y}_e)^2 = .05)$; well-structured tasks: $(r(y_e, y_s \cdot \hat{y}_e)^2 = .03)$). Therefore, Hypothesis 2 holds true regardless of the level of task-relevant expertise. The maximum percentage of unexplained model variance picked up by respondents is 58% for Major Label Experts and 55% for Independent Label Experts.

Hypothesis 3 postulates that an increase in the number of aggregated judgments will yield a marginally diminishing improvement in forecasting accuracy. This is fully supported by the music industry data. To test Hypothesis 3, we analyzed the impact of four different methods of individual judgment aggregation on the relative importance of model and judges: equal unit weighting, outperformance, optimal weighting and regression (Fig. 3). In our comparisons, we focused on groups of 3, 6 and 12 individual respondents for both well- and ill-structured forecasting tasks. In all scenarios, aggregated judgments account for at least 65% of the combination weights.

For more refined insights into the effect of group sizes, we converted the standardized beta coefficients of each model-multiple judge combination into relative weights for each possible group size when relying on equal unit weights. Fig. 4 displays the results of these analyses. In line with previous findings on the aggregation

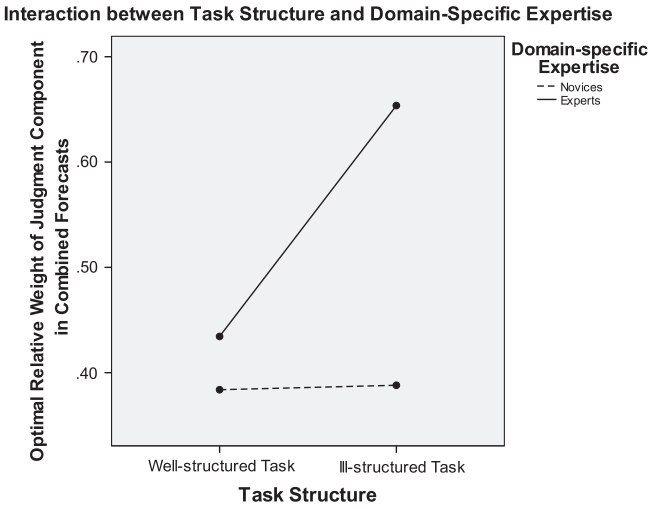

Fig. 1. 2 × 2 Factorial ANOVA: Task structure, domain-specific expertise and the relative importance of human judgment.

of individual judgments (Hogarth, 1978), the benefit arising from the inclusion of one additional expert to the aggregated judgments decreases marginally, a result that holds true regardless of the type of aggregation method used, levels of expertise, and task structure. While the improvement in accuracy when adding one additional expert marginally diminishes, our results show that the largest incremental improvement occurs between the first and second added judgment (Major Label Experts: $\Delta = 0.09$; Independent Label Experts: $\Delta = 0.08$), supporting Libby and Blashfield's (1978) earlier findings. This result suggests that the largest marginal improvement in accuracy occurs with up to three individual judgments. Diverging levels of expertise also seem to translate as only small differences in predictive performance (Major Label Experts: mean $\Delta = .05$; Independent Label Experts: mean $\Delta = .07$). Adding more experts to model-multiple judge combinations results in higher relative weights assigned to human judgment, which validates Hypothesis 3.

Discussion and conclusions

Our study provides clear answers to the fundamental debate on the degree to which forecasts should rely on linear models or human judges. In the context of the music industry, Major Label Experts achieved the highest predictive accuracy, picked up the highest percentage of unexplained model variance, and exhibited the greatest validity of expertise among all forecasters. As expected, linear models outperformed respondents least familiar with the predictive task at hand (Novices). The 2% of unexplained model variance picked up by Novices regardless of task structure (as shown in the third and sixth lines of Table 1) can be attributed to luck. Within their specific social territory (that is, the specific social context that controls how they relate to music, and where they subconsciously interpret the cognitive associations stirred up by music as particular emotions), Novices may also have been conditioned to recognize certain musical cues as characteristic of a hit song (Witchel, 2010). Experts, who contribute to defining the standards in the social territory of pop music, are also better at identifying musical cues (e.g., notes, harmony, pace, instruments, and lyrics) that make a "hit" and at predicting the success potential of new artists.

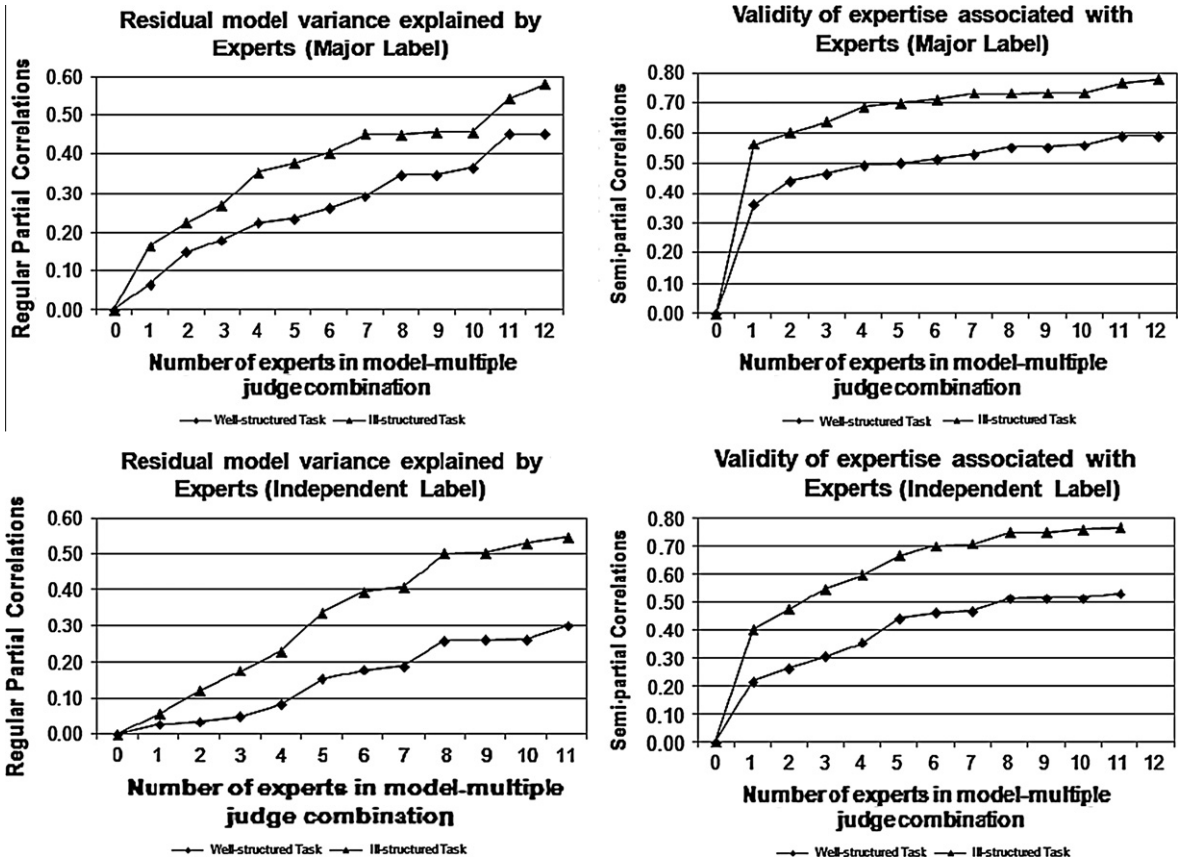

Fig. 2. Mean values of regular partial and semi-partial correlations.

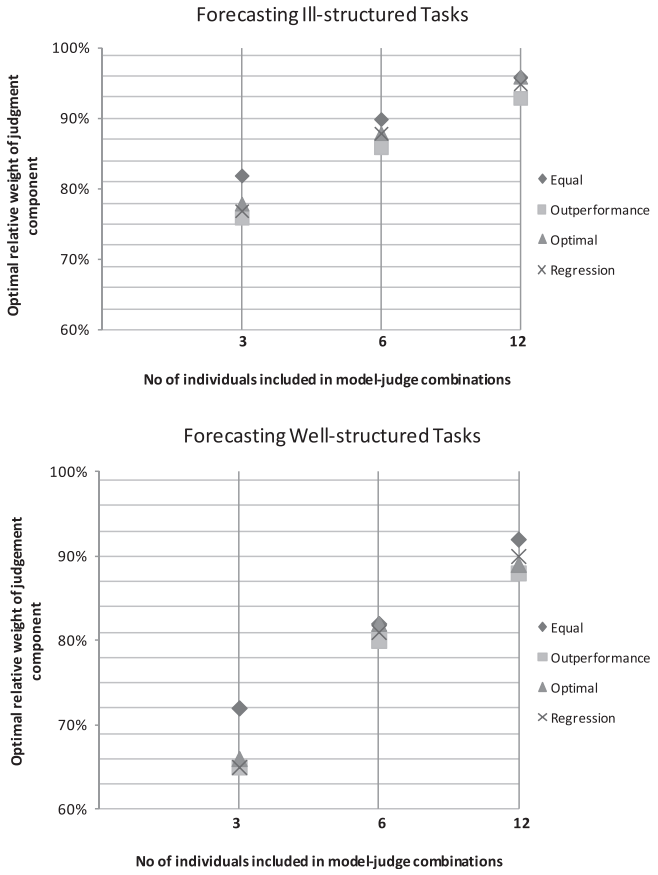

Fig. 3. Effect of aggregation method on the optimal relative weight of judgmental component (3, 6 and 12 Individual judges).

Individual judgments prove most powerful when Major Label Experts assess the chart entry position of newcomers. When making predictions in ill-structured tasks conditions, Major Label Experts picked up 2.5 times more unexplained model variance than Independent Label Experts, and 8 times more than Novices. These individual-level effects are consistent across the number of respondents included in each model-multiple judge combination, regardless of the method used for aggregating these respondents' predictions. An increase in the number of aggregated judgments also decreases the relative importance of linear models, regardless of forecasting task structuredness and of the level of domain-specific expertise involved.

We ran additional tests to control for robustness and to test alternative explanations for our findings. First, we investigated whether the delay between respondents' judgments and the actual release date of the pop music single influenced our results. The answer to this question is 'no', as having more timely information available did not improve respondents' predictive accuracy. Second, we tested alternative measures of domain expertise, such as by sorting respondents according to industry tenure, past employment in major or independent labels, and past hit-rate. While these modifications resulted in slight variations in respondents' ability to pick up nonlinearities when predicting the success of newcomers, our main findings were robust. Third, we tested for the effect of inside knowledge that experts may possess when making predictions about artists signed by their own labels (for instance, in terms of upcoming promotional activities or access to market research data). Again, the resulting slight improvement in predictive accuracy did not change our conclusions in a statistically significant way.

Ultimately, we offer the following conceptual and empirical contributions to the judgmental forecasting literature. First, our results extend Blattberg and Hoch's (1990) seminal study of model-manager combinations by offering novel insights into the role of individual and aggregated judgments. In particular, we use a unique empirical real world setting to provide one of the rare tests of the robustness of their proposed 50/50 split between model output and human judgment. We demonstrate that the proposed split may be sensitive to changes in individual respondents' levels of domain-specific expertise. By relying on the principle of task familiarity (Hammond, 1996) to assess differing levels of expertise, we show that domain-specific expertise has a positive effect on the relative importance of the judgmental component. This result is confirmed by our 2×2 factorial ANOVA, which identifies a main effect for the level of domain-specific expertise associated with the human judgment component in model-judgment combinations.

Second, our research reconciles contradictory findings relative to the predictive performance of linear models versus human judges (Lawrence, Goodwin, O'Connor, & Önkal, 2006; Makridakis et al., 1993). Studies largely conducted in laboratory settings suggest that models outperform experts, whereas other studies typically conducted in natural settings claim the contrary. In line with Hypotheses 1a-c, our findings support the idea that this contradiction may be due to the moderating effect of task structure. Artificial forecasting tasks used in laboratory studies are likely to be based on less variable data structures than naturalistic prediction environments. It is possible that the necessity of controlling the task in an experiment produces biased forecasting environments, in which linear models are likely to outperform human judges precisely because high forecasting accuracy can be achieved in the laboratory with less contextual knowledge. In naturalistic settings, what type of information will be useful for building effective linear models and for improving forecasting accuracy is often unclear. In ill-structured contexts, human judges are therefore likely to outperform linear models when they can rely on large amounts of domain specific expertise.

Third, our study is novel in explicitly investigating the interaction between task structure and domain-specific expertise in the context of combined model-judge forecasts. Comparing and contrasting differences between well-structured and ill-structured forecasting tasks demonstrates that the outcome effectiveness of judgments critically hinges upon the properties of the task. The resulting interaction graph indicates that when considering well-structured tasks, the model forecast takes precedence, regardless of domain-specific expertise. In contrast, when facing ill-structured forecasting tasks, an increase in domain-specific expertise shifts the optimal model-judge split towards a heavier reliance on human judgment. The empirical findings outline the conditions in which it is desirable to employ expert judges when generating predictions, and both reinforce and mitigate the evidence-based management approach.

In the wake of a natural science of organizations (Barnard, 1938), evidence-based management aims to help organizations perform better by moving professional decisions away from personal preference and unsystematic experience. Instead, evidence-based management advocates an active use of the best available sources of systematic knowledge and scientific evidence of the cause-effect principles underlying human behavior and organizational actions (Pfeffer & Sutton, 2006b; Rousseau, 2006).

Our results reinforce the evidence-based management approach, inasmuch as they resonate with its core idea that decisions should be reached "through the consciencious, explicit, and judicious use of four sources of information: practitioner expertise and judgment, evidence from the local context, a critical evaluation of the best available research evidence, and the perspectives of

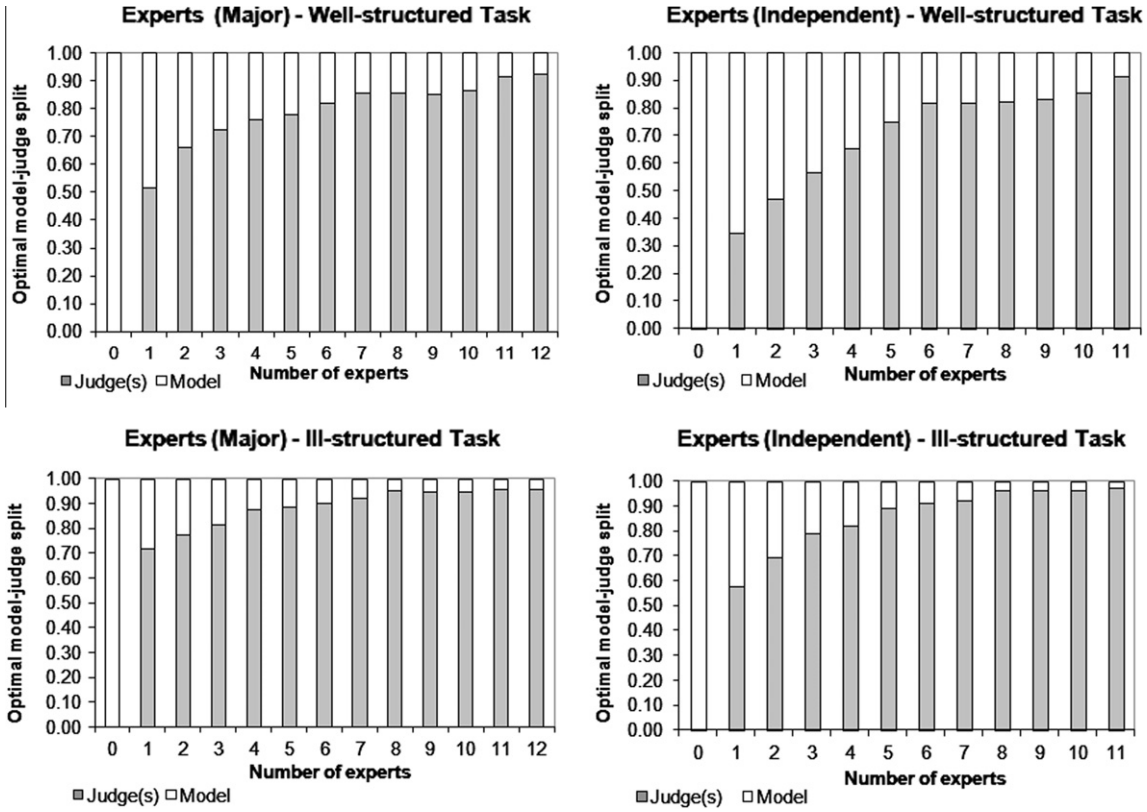

Fig. 4. Optimal relative weights in combined model-multiple expert forecasts.

those people who might be affected by the decision" (Briner, Denyer, & Rousseau, 2009, p. 19). Our results mitigate, however, the core assumption of some evidence-based management authors that linear models are inherently superior to human judgment (Lawler, 2007), and that relying too much on the latter may be dangerous (Briner & Rousseau, 2011). Although we support the general efficacy of linear models, our study presents a counterpoint to some of the most discussed real world examples of evidence-based management, including the Moneyball narrative mentioned in the introduction to this article. Rather than advocating eschewing human judgments altogether and making player recruitment decisions solely on the basis of mathematical models (Lewis, 2004), our findings suggest that, at least under some conditions, models can be improved by incorporating human judgment. They therefore provide an illustration of how evidence-based management need not be devoid of human judgment but, rather, can benefit from the systematically and statistically justified inclusion of it. As such, they elude the danger of: "privileging research evidence over other forms of evidence [...] and insights from other sources, especially professional experience" (Briner et al., 2009, p. 20).

Fourth, the study offers empirical insights into the value of model-multiple judge combinations associated with mechanically aggregated group judgments. When considering ill-structured forecasting tasks, the increased value of aggregated judgments appears related to an improvement in the collective ability to interpret nonlinearities in the task environment. Conversely, when considering well-structured tasks, superior forecasting performance is likely to result from a more efficient interpretation of linear relationships in the task ecology. Our test of Hypothesis 3 indicates that when mechanically adding human judgments to model-multiple judgment combinations, the largest marginal improvement in predictive accuracy occurs with up to three individual judges. The improvement in predictive accuracy brought about by the aggregation of multiple judgments also marginally diminishes as the number of human judges involved in model-multiple judgment combinations increases.

We fully acknowledge that our conclusions have been derived from only one empirical context, and we therefore believe that this study unlocks a number of additional research opportunities. Studies of aggregated judgments could use similar regression models to capture performance differentials in a wider range of forecasting contexts. Future research could also investigate combined forecasts involving models that capture both linear and nonlinear relationships among informational cues in the environment. Moreover, a comparative study of managers in stable versus unstable industries and environments could further illustrate the conditions under which individual, aggregated, or collective judgment is to be preferred.

In practice, marketing and A&R managers in the music industry could use our findings as guidelines for aggregating statistical model predictions and managerial judgments of the success potential of artists, singles and albums, particularly in the context of a severe crisis brought about by illegal file sharing and internet piracy. By distinguishing between ill- and well-structured forecasting tasks, our study delineates the circumstances under which higher reliance on linear models or judgmental predictions are likely to improve forecasting accuracy. In line with evidence-based management recommendations (Rousseau, 2006), our results provide an encouragement to A&R managers to move away from predictions based exclusively on "gut feeling" and to complement their judgment with the scientific evidence provided by linear models[11] in order to reach optimal decisions in selecting artists to

[11] Note however, in response to Learmonth and Hardings's (2006) critique of evidence-based management, that linear models only represent a particular, dominant approach to constructing evidence in the music industry. By advocating the use of linear models, we do not aim to disregard or render other alternatives to evidence construction invisible, or to treat the evidence they provide as universal.

sign. Particularly when dealing with well-structured tasks, demanding such evidence and concentrating on "facts" rather than "brags" will allow them to move away from guesswork, fear, belief, or hope and make better, more logical decisions (Pfeffer & Sutton, 2006a). We similarly encourage A&R managers and other decision-makers to act on the basis of the best knowledge available at a given time, whilst remaining open to learning, seeing and doing things differently, because in the future, that knowledge may be found to be incorrect (Pfeffer & Sutton, 2007).

More generally, our findings may also prove valuable in other blockbuster-type decision environments, where the prediction of product success and the creation of optimal product portfolios are equally ambiguous and dependent on the correct identification of extremely profitable "outliers." In particular, our conclusions could help professionals in other cultural industries, such as publishing and cinema, forecast the development, distribution and sales of new books and movies and identify potential and established star authors, directors or actors.

Reflecting on the examples mentioned in the introduction to this article, it appears obvious that the IBM Watson supercomputer is far more advanced than the simple linear model underlying this study. Nevertheless, replacing human judgment with a medical diagnosis generated by a machine will inevitably involve important ethical concerns, so that the debate, which is at the heart of our study, of when and how to rely on Watson is of highest importance. In the same way, relying exclusively on trading algorithms for complex financial decisions seems ill-advised. Finally, in light of our empirical findings, the disappointing outcome of Oakland Athletics general manager Billy Beane's purely mathematical approach to selecting team members isn't too surprising. Forecasting in the sports industry is highly uncertain, because key variables (competitors, managers, funding, etc.) constantly change and human performance is highly unreliable. Fluctuations in players' performance from one season to the next therefore cause substantial ambiguity in the input data of any model which uses human performance as its sole parameters. Taking this into account, in the broadest sense our study offers relatively simple guidance: (1) always combine model and judge(s), (2) Always assess the structure of a task, and (3) The more ill-structured the task appears to be, the more likely it is that expert diagnosis will be preferable.

Acknowledgments

The authors would like to thank Martin Kilduff, Eugene Sadler-Smith, Stefan Scholtes, Lisa Bolton, Madan Pillutla, Mark de Rond and the participants of the Management Science Seminar at University of Cambridge Judge Business School for their helpful feedback and insightful discussions. They are also grateful to Professor William Bottom and the three anonymous reviewers from OBHDP for their invaluable comments during the reviewing process. Lastly, the authors appreciated the generous financial support of the Friedrich Naumann Foundation and the Cambridge European Trust (M. Seifert).

References

Alexander, J. (1995). Refining the degree of earnings surprise: A comparison of statistical and analysts' forecasts. *Financial Review, 30*(3), 469–506.
Armstrong, S. (1983). Relative accuracy of judgmental and extrapolative methods in forecasting annual earnings. *Journal of Forecasting, 2*, 437–447.
Ashton, R. H. (1986). Combining the judgments of experts: How many and which ones? *Organizational Behavior and Human Decision Processes, 38*, 405–414.
Barnard, C. I. (1938). *The functions of the executive*. Cambridge, MA: Harvard University Press.
Bates, J., & Granger, C. (1969). The combination of forecasts. *Operations Research Quarterly, 20*, 451–468.
Blattberg, R. C., & Hoch, S. J. (1990). Database models and managerial intuition: 50% model + 50% manager. *Management Science, 36*(8), 887–899.
Briner, R. B., Denyer, D., & Rousseau, D. M. (2009). Evidence-based management: Concept cleanup time? *Academy of Management Perspectives* (November), 19–32.
Briner, R. B., & Rousseau, D. M. (2011). Evidence-based I–O psychology: Not there yet. *Industrial and Organizational Psychology, 4*, 3–22.
Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley, CA: The University of California Press.
Bunn, D. W. (1975). A Bayesian approach to the linear combination of forecasts. *Operational Research Quarterly, 26*(2), 325–329.
Bunn, D. W. (1985). Statistical efficiency on the linear combination of forecasts. *International Journal of Forecasting, 1*, 15–163.
Bunn, D. W. (1987). Expert use of forecasts: Bootstrapping and linear models". In G. Wright & P. Ayton (Eds.), *Judgmental forecasting* (pp. 229–241). New York: Wiley.
Camerer, C., & Weber, M. (1992). Recent developments in modeling preferences: Uncertainty and ambiguity. *Journal of Risk and Uncertainty, 5*, 325–370.
Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5*, 559–583.
Cohen, R. (1993). *France and Spain impose quotas*. New York Times, 15 December: Section C:15.
Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis of the behavioral sciences*. Hillsdale, NJ: Erlbaum.
Cooksey, R. (1996). *Judgment analysis: Theory, methods and applications*. NY: Academic Press.
Cowen, T., & Grier, K (2011). *The economics of moneyball: Do the principles really work anymore?* <http://www.grantland.com/story/_/id/7328539/the-economics-moneyball> (7 December).
Dawes, R., Faust, D., & Meehl, P. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 68–74.
Deng, P. S. (1996). Using case-based reasoning approach to the support of ill-structured decisions. *European Journal of Operational Research, 93*(3), 511–521.
Dunwoody, P. T., Haarbauer, E., Mahan, P., Marino, C. J., & Tang, C. C. (2000). Cognitive adaptation and its consequences: A test of cognitive continuum theory. *Journal of Behavioral Decision-making, 13*, 35–54.
Einhorn, H. J. (1974). Cue definition and residual judgment. *Organizational Behavior and Human Decision Processes, 12*, 30–49.
Einhorn, H. J., & Hogarth, R. M. (1975). Unit weighting schemes for decision-making. *Organizational Behavior and Human Decision Processes, 13*, 171–192.
Einhorn, H. J., Hogarth, R. M., & Klempner, E. (1977). Quality of group judgment. *Psychological Bulletin, 84*(1), 158–172.
Fellner, W. (1961). Distortion of subjective probabilities as a reaction to uncertainty. *Quarterly Journal of Economics, 75*, 670–694.
Feltovich, P., Prietula, M., & Ericsson, A. (2006). Studies of expertise from psychological perspectives. In A. Ericsson, N. Charness, P. Feltovich, & R. Hoffman (Eds.), *Expertise and expert performance* (pp. 41–69). New York: Cambridge University Press.
Frisch, D., & Baron, J. (1988). Ambiguity and rationality. *Journal of Behavioral Decision Making, 1*, 149–157.
Gapper, J. (2008). Guy Hands versus musical rent-seekers. *Financial Times*.
Goldberg, L. R. (1970). Man versus model of man: A rationale, plus some evidence for a method of improving on clinical inferences. *Psychological Bulletin, 73*, 422–432.
Goodwin, P. (2000). Correct or combine? Mechanically integrating judgmental forecasts with statistical methods. *International Journal of Forecasting, 16*(2), 261–275.
Hammond, K. R. (1988). Judgment and decision-making in dynamic tasks. *Information and Decision Technologies, 14*(2), 3–14.
Hammond, K. R. (1996). *Human judgment and social policy*. New York: Oxford University Press.
Hitquarters (2008). *A&R network*. <http://www.hitquarters.com> Retrieved 25.05.08.
Hoch, S. J., & Schkade, D. A. (1996). A psychological approach to decision support systems. *Management Science, 42*(1), 51–64.
Hoch, Stephen J. (1987). Perceived consensus and predictive accuracy: The pros and cons of projection. *Journal of Personality and Social Psychology, 53*(2), 221–234.
Hoch, Stephen J. (1988). Who do we know: Predicting the interests and opinions of the american consumer. *Journal of Consumer Research, 15*(3), 315–324.
Hogarth, R. M. (1978). A note on aggregating opinions. *Organizational Behavior and Human Decision Processes, 21*, 40–46.
Hogarth, R. M. (2001). *Educating intuition*. Chicago: Chicago University Press.
IFPI (2011). *Statistics yearbook on global recorded music market*. <http://www.ifpi.org> Retrieved 14,.06.08.
Kahneman, D., & Tversky, A. (1984). Choices, values and frames. *American Psychologist, 39*, 341–350.
Klein, G. (2003). *Intuition at work*. New York: Doubleday.
Kleinmuntz, D. (1990). Why we still use our heads instead of formulas: Toward an integrative approach. *Psychological Bulletin, 107*(3), 296–310.
Lawler, E. E. (2007). Why HR practices are not evidence-based. *Academy of Management Journal, 50*(5), 1033–1036.
Lawrence, M., Edmundson, R., & O'Connor, J. F. (1986). The accuracy of combining judgmental and statistical forecasts. *Management Science, 32*, 1521–1532.
Learmonth, M., & Harding, N. (2006). Evidence-based management: The very idea. *Public Administration, 84*(2), 245–266.
Lawrence, M., Goodwin, P., O'Connor, M., & Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting, 22*, 493–518.

Lewis, M. (2004). *Moneyball: The art of winning an unfair game*. New York, London: W.W. Norton & Company.
Libby, R., & Blashfield, R. (1978). Performance of a composite as a function of the number of judges. *Organizational Behavior and Human Decision Processes, 21*, 121–129.
Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, K., et al. (1993). The M2 competition: A real time judgmentally based forecasting study. *International Journal of Forecasting, 9*, 5–22.
McClish, D. K., & Powell, S. H. (1989). How well can physicians estimate mortality in a medical intensive care unit? *Medical Decision Making, 9*, 125–132.
Meehl, P. (1996). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. Northvale, NJ: Jason Aronson.
de Menezes, L. M., Bunn, D. W., & Taylor, J. W. (2000). Review of guidelines for the use of combined forecasts. *European Journal of Operational Research, 120*(1), 190–204.
Mentzer, J. T., & Moon, M. A. (2005). *Sales forecasting management* (2nd ed.). London, UK: Sage Publications.
Moe, W. W., & Fader, P. S. (2001). Modeling hedonic portfolio products: A joint segmentation analysis of music compact disc sales. *Journal of Marketing Research, 38*, 376–385.
Peters, J. T., Hammond, K. R., & Summers, D. A. (1974). A note on intuitive vs. analytic thinking. *Organizational Behavior and Human Decision Processes, 12*(2), 125–131.
Pfeffer, J., & Sutton, R. I. (2006a). Evidence-based management. *Harvard Business Review* (January), 1–12.
Pfeffer, J., & Sutton, R. I. (2006b). *Hard facts, dangerous half-truths, and total nonsense: Profiting from evidence-based management*. Boston, MA: Harvard Business School Press.
Pfeffer, J., & Sutton, R. I. (2007). Suppose we took evidence-based management seriously: Implications for reading and writing management. *Academy of Management Learning & Education, 6*(1), 153–155.
Rousseau, D. M. (2006). Is there such thing as "evidence-based management"? *Academy of Management Review, 31*(2), 256–269.
Sanders, N., & Ritzman, L. (1995). Bringing judgment into combination forecasts. *Journal of Operations Management, 13*, 311–321.
Schoemaker, P. J. H. (1993). Strategic decisions in organizations: Rational and behavioral views. *Journal of Management Studies, 30*(1), 108–129.
Shapira, Z. (2002). *Organizational decision making*. New York: Cambridge University Press.
Silverstein, A. B. (1987). Equal weighting vs. differential weighting of subtest scores on short forms of Wechsler's intelligence scales. *Journal of Clinical Psychology, 43*(6), 714–720.
Simon, H. A. (1977). *The new science of management decision*. Englewood Cliffs, NJ: Prentice-Hall.
Spence, M. T., & Brucks, M. (1997). The moderating effects of problem characteristics on experts' and novices' judgments. *Journal of Marketing Research (JMR), 34*(2), 233–247.
Steinmann, D. (1976). The effects of cognitive feedback and task complexity in multiple-cue probability learning. *Organizational Behavior and Human Performance, 15*, 168–179.
Stewart, T. R., Roebber, P. J., & Bosart, L. F. (1997). The importance of the task in analyzing expert judgment. *Organizational Behavior & Human Decision Processes, 69*(3), 205–219.
Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*. New York: Random House Inc.
Vogel, H. L. (2004). *Entertainment industry economics – A guide for financial analysts* (5th ed.). Cambridge: Cambridge University Press.
Witchel, H. (2010). *You are what you hear: how music and territory make us who we are*. London: Algora Publishing.
Wood, R. E. (1986). Task complexity: Definition of the construct. *Organizational Behavior and Human Decision Processes, 37*, 60–82.
Yaniv, I., & Hogarth, R. M. (1993). Judgmental versus statistical prediction: information asymmetry and combination rules. *Psychological Science* (Wiley-Blackwell) 4(1), 58–62.
Zufryden, F. S. (1996). Linking advertising to box office performance of new film releases – a marketing planning model. *Journal of Advertising Research* (July–August), 29–41.