Received: 21 March 2018 | Revised: 12 February 2019 | Accepted: 13 February 2019

DOI: 10.1002/bdm.2123

RESEARCH ARTICLE

WILEY

Optimal cue aggregation in the absence of criterion knowledge

Wenjia Joyce Zhao[1] | Clintin P. Davis-Stober[2] | Sudeep Bhatia[1]

[1] Department of Psychology, University of Pennsylvania, Philadelphia, Pennsylvania

[2] Department of Psychological Sciences, University of Missouri, Columbia, Missouri

Correspondence
Wenjia Joyce Zhao, Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104.
Email: zhaowenj@sas.upenn.edu

Abstract

The study of multi-cue judgment investigates how decision makers aggregate cues to predict the value of a criterion variable. We consider a multi-cue judgment task in which decision makers have prior knowledge of inter-cue relationships but are ignorant of how the cues correlate with the criterion. In this setting, a naive judgment strategy prescribes weighting the cues equally. Although many participants are well described via an equal weighting scheme, we find that a substantial minority of participants make predictions consistent with a weighting scheme based on a low-dimensional projection of the cue space that optimally takes into account inter-cue correlations. The use of such a weighting scheme is consistent with minimizing maximal error in prediction when the cue-criterion relationships are unknown.

KEYWORDS

cue integration, dimensionality reduction, improper linear models, judgment and decision making

1 | INTRODUCTION

Effective judgment and decision making involves the aggregation of multiple cues, or pieces of information, to evaluate a criterion variable. For example, an individual may receive advice from two or more friends regarding a financial investment and aggregate this advice to make a prediction on the investment's return. Alternatively, a supervisor may have to choose between job candidates, each with multiple attributes and qualifications, and must somehow use this information to determine the overall quality of the candidates. Understanding the mental processes used by people to aggregate multiple cues and make judgments is an important topic of research in psychology (for reviews, see Gigerenzer & Gaissmaier, 2011; Oppenheimer & Kelso, 2015; Weber & Johnson, 2009; also see Bhatia, 2018; Juslin, Karlsson, & Olsson, 2008; Marewski & Mehlhorn, 2011; Newell & Bröder, 2008). Traditionally, researchers employ a linear modeling approach when evaluating judgment and decision making behavior in both normative and descriptive contexts. This approach models decision makers as computing the value of a criterion using a weighted average of cues, with the weights being proportional to the observed relationship between the cues and the criterion (e.g., Brunswik, 1952; Keeney & Raiffa, 1993). In this article, we consider a ubiquitous decision making environment that is seldom studied within the literature. Specifically, we consider environments where decision makers have information about how cues correlate with one another but have no information about how cues correlate with the criterion variable. For example, individuals using the advice of their friends to judge investments may not have previously observed how well their friends actually predict the performance of such investments. Likewise, individuals evaluating job candidates for novel or unconventional jobs may have never observed the value of different candidate attributes in the context of such jobs. In these situations, decision makers may have detailed knowledge about the relationship between the cues (e.g., how often their friends agree with each other or how frequently job candidate attributes co-occur) but have no way to assign weights to the cues in accordance with the standard linear model (where weights depend on the cues' relationship with the criterion).

For such cases, "improper linear models" (Dawes, 1979) can be applied to evaluate various decision strategies within a linear modeling context. The major deviation from standard methods is that the weights for the predictor cues are determined without any information about the relationships between cues and the criterion variable. These models involve a fixed weighting scheme that assigns a priori relationships among the cue weights. For example, an equal weights improper linear model would give each cue the same weight (Dawes, 1979; Dawes, Faust, & Meehl, 1989; Fishburn, 1974; Gigerenzer,

Todd, ABC Research Group, 1999; Payne, Bettman, & Johnson, 1993), whereas a simple lexicographic model would assign weights in such a way that a single cue (or a specific ordering of cues) would drive the decision, uninfluenced by the values of the other cues (Czerlinski, Gigerenzer, & Goldstein, 1999; Martignon & Hoffrage, 2002). Dana and Davis-Stober (2016) provide a comprehensive discussion of how improper linear models can be used to model and evaluate a large class of decision strategies and heuristics. By simplifying cue weighting, improper linear models are less cognitively demanding for decision makers. They are also more robust, less vulnerable to overfitting, and, for this reason, have been shown to outperform proper linear models in many situations, ranging from graduate student admission to clinical predictions (Davis-Stober, 2011; Dawes, 1979; Gigerenzer & Gaissmaier, 2011).

For multi-cue judgment with known inter-cue relationships, but unknown cue-criterion relationships, the key questions of interest are the following: Which improper weighting scheme should decision makers use and which schemes do decision makers use? The former question has been addressed by Davis-Stober, Dana, and Budescu (2010a, 2010b). Davis-Stober et al. propose that any possible weighting scheme, $\boldsymbol{\beta}$, can be assessed with regard to how far it deviates from the true (population) weight vector, $\boldsymbol{\beta}^*$. When the cue-criterion relationships are unknown, $\boldsymbol{\beta}^*$ is also unknown. For the environment described above, where no information is available about the relationship between the predictors and the dependent variable, one can still consider an optimal $\boldsymbol{\beta}$ with regard to minimizing maximum risk, where risk is defined as the expected sum of squared errors between the elements of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}^*$. By this standard, the best improper linear weighting scheme is one that is proportional to the eigenvector corresponding to the first (i.e., largest) eigenvalue of the cue correlation matrix (see Davis-Stober et al., 2010a, for details, and Lord & Novick, 1968, for an early discussion). Intuitively, this weighting scheme utilizes the best one-dimensional projection of the predictor cue correlation space. We will refer to this weighting scheme as $\boldsymbol{\beta}_{\boldsymbol{EV1}}$. Here, *EV1* in the subscript refers to the use of the eigenvector corresponding to the first (i.e., largest) eigenvalue. If a decision maker uses such a weighting scheme, then her behavior is consistent with minimizing maximum risk. Said simply, she is using the information that is available to her (the relationship among cues) to minimize her worst-case errors in cue weighting. Note that the use of this eigenvector is optimal only in the sense that it minimizes maximal risk. Without information about criterion values, other improper linear models, such as the equal weight strategy, can also provide robust predictions.

To clarify, there are two interpretations at play when considering an individual using the $\boldsymbol{\beta}_{\boldsymbol{EV1}}$ weighting scheme. First, in a fundamental way, the primary eigenvector, $\boldsymbol{\beta}_{\boldsymbol{EV1}}$, is the linear combination that accounts for the maximal variance possible in a one-dimensional representation of the covariance matrix of predictor cues. In other words, weighting the predictor cues in this way accounts for more of the total variance (among the predictor cues) compared with any other weighting scheme. This is irrespective of any relationship with the criterion variable and is extremely general. One interpretation is that, upon learning the covariance matrix of the cues, the decision maker weights the cues using $\boldsymbol{\beta}_{\boldsymbol{EV1}}$ to make choices as it is the most faithful one-dimensional representation of the covariance matrix. Second, if we consider the relationship with the criterion variable, then $\boldsymbol{\beta}_{\boldsymbol{EV1}}$ is minimax compared with any other weighting scheme. This means that $\boldsymbol{\beta}_{\boldsymbol{EV1}}$ minimizes maximal risk, that is, the "worst-case" relationship between the criterion variable and predictor cues is the smallest compared with any other weighting scheme. This is highly relevant to our experimental design as the decision maker has *no information* about the relationship between the criterion variable and the predictor cues. Thus, minimizing worst-case outcomes is one of the (very few) options available to the decision maker. There simply are not many other objective functions one could consider.

Normative solutions aside, which weighting scheme do decision makers actually use when integrating multiple cues with unknown cue-criterion relationships? A first guess involves an equal weights model: without knowing that cues are more related to the criterion than others, it seems conceivable that decision makers assign the same weights to all the available cues. This corresponds to a type of ignorance prior, which has been shown to work well in many decision contexts (Armstrong, 2001; Clemen, 1989; Fischer & Harvey, 1999). Of course, the equal weights heuristic has also been shown to be a good descriptor of human behavior (see, e.g., Gigerenzer & Gaissmaier, 2011).

In contrast, some prior work suggests that decision makers are sensitive to inter-cue relationships. Particularly, confidence in estimates obtained through cue aggregation is higher when there is more agreement between the cues (Bonaccio & Dalal, 2006; Sniezek & Buckley, 1995; Tversky & Kahneman, 1973). Additionally, the learning of cue-criterion relationships (and subsequently accuracy in predicting the criterion) is sensitive to the structure of inter-cue correlation (Armelius & Armelius, 1976; Klayman, 1988; but also see Maines, 1996; Soll, 1999). Finally, in settings where cues themselves depend on a shared set of information sources, individuals are able to take into account the shared information structure and control for the resulting cue correlations, when predicting criterion values (Broomell & Budescu, 2009; Budescu & Rantilla, 2000; Budescu, Rantilla, Yu, & Karelitz, 2003).

Although the $\boldsymbol{\beta}_{\boldsymbol{EV1}}$ weighting scheme appears desirable from a normative perspective, its use also coincides with established findings regarding semantic representation. Decision makers with prior experience with a set of object features are able to learn mental representations of the feature space. These representations, in many settings, correspond to low-dimensional projections of the decision makers' experiences with the features. Such projections can be approximated by a principle components analysis on the feature-correlation matrix, or equivalently, a singular value decomposition on the matrix of feature-object co-occurrence. Indeed, such a decomposition is a key component of numerous existing approaches to modeling semantic representation, including latent semantic analysis (Landauer & Dumais, 1997), multidimensional scaling (Kruskal & Wish, 1978), and neural network models of semantic memory (Saxe, McClelland, & Ganguli, 2013). Interestingly, such a decomposition resembles the $\boldsymbol{\beta}_{\boldsymbol{EV1}}$ model when only the first dimension of the decomposed representation is used. Thus, it seems that the $\boldsymbol{\beta}_{\boldsymbol{EV1}}$ weighting scheme may not only

be optimal from a statistical perspective but may also parallel the way in which semantic knowledge is learnt and represented in other domains in psychology.

In sum, it seems that judges do not ignore cue correlations and that they can even use correlation information in an efficient manner. This in turn suggests that instead of using an equal weighting scheme, in the absence of cue-criterion knowledge, individuals may be able to safeguard against worst-case outcomes by utilizing cue correlation information as recommended by the β_{EV1} weighting scheme. The use of such a weighting scheme would imply that behavior is "rational," in that it appropriately responds to environmental statistics to optimize relevant decision goals (Anderson, 1990; Gigerenzer & Gaissmaier, 2011; Griffiths & Tenenbaum, 2016; Oaksford & Chater, 2007; Tenenbaum, Griffiths, & Kemp, 2006).

2 | IMPROPER LINEAR MODELS IN THE ABSENCE OF CUE-CRITERION KNOWLEDGE

We examine improper linear models that decision makers use to integrate cues, in a situation with unknown cue-criterion correlations but known inter-cue correlations. As is suggested by the normative β_{EV1} weighting scheme, in the absence of the cue-criterion knowledge, decision makers can perform better than simply averaging the cue values by applying cue correlation knowledge and weighting cues in a fashion that is consistent with the primary eigenvector weighting.

Improper linear models are applied within the context of a standard regression framework with dependent variable $\boldsymbol{y}$ and (standardized) predictor cues $\boldsymbol{X}$, that is, $\boldsymbol{y = X\,\beta + e}$, where $\boldsymbol{e}$ is a normally distributed random variable with mean equal to 0. In a typical regression framework, data are used to estimate $\boldsymbol{\beta}$, often via ordinary least squares. Improper models differ in that the relationships between the values of $\boldsymbol{\beta}$ are not estimated via data. While one could use data to set the scale in improper linear models (as in Davis-Stober et al., 2010a) how much larger one weight in $\boldsymbol{\beta}$ is than another is not allowed to change. That is, the relative weights assigned to different cues are not allowed to change. For example, one could prescribe an improper linear model such that the weights corresponding to the first two cues will be twice as large as the remaining cues. The basic intuition of why improper linear models perform well can be described via expected squared error, that is, risk. Risk can be written as the sum of an estimator's bias and its variance. Improper linear models are clearly biased, in that, on average, they do not equal the true weights, $\boldsymbol{\beta}^*$. On the other hand, depending upon the setup, they have little to no variance. Given noisy environments with limited data, improper linear models can outperform standard estimation techniques in predicting future observations by reducing variance at the expense of bias (see Davis-Stober, 2011, for a comprehensive discussion).

The primary eigenvector weighting, β_{EV1}, can be shown, in appropriate settings, to approximate other well-known improper linear models. Consider, for example, the cue correlation matrix in Figure 1a. Here, Cue 1 is highly correlated with the remaining three cues with a coefficient of 0.6, and all the remaining cues (Cues 2–4) are weakly correlated internally with a coefficient of 0.05. For ease of presentation, we will write our weighting vectors with normed length equal to 1. For this cue correlation matrix, we obtain β_{EV1} = [0.69, 0.42, 0.42, 0.42], implying that the normative weighting scheme overweights Cue 1 and underweights Cues 2–4. Said differently, a decision maker using a linear rule proportional to this vector will place just over 1.5 times as much weight on the first cue compared with the remaining cues (which will be equally weighted). This can be seen as a milder form of a lexicographic decision rule, which disproportionately values one of the cues while ignoring the others. In contrast, consider the cue correlation matrix in Figure 1b, where all the cues are moderately correlated with each other with a coefficient of 0.4. Here, the β_{EV1} weighting scheme weighs each cue equally, with β_{EV1} = [0.5,0.5,0.5,0.5], thus any linear rule proportional to this vector is using an equal weights weighting scheme.

(a)

	C1	C2	C3	C4
C1	1			
C2	.6	1		
C3	.6	.05	1	
C4	.6	.05	.05	1

(b)

	C1	C2	C3	C4
C1	1			
C2	.4	1		
C3	.4	.4	1	
C4	.4	.4	.4	1

(c)

	C1	C2	C3	C4
C1	1			
C2	.8	1		
C3	.1	.1	1	
C4	.1	.1	.4	1

FIGURE 1 (a–c) Three cue correlation matrices used in this paper. Figure 1a shows the cue correlation matrix for Study 1 and the treatment condition of Study 2. Figure 1b shows the cue correlation matrix for the control condition of Study 2. Figure 1c shows the cue correlation matrix for Study 3

3 | GENERAL METHOD

Our goal was to investigate the plausibility of the β_{EV1} weighting scheme and to compare its ability to predict participant judgments with alternate improper linear models such as the equal weights rule and the lexicographic rule. To appropriately test these models, we examined settings in which participants had prior knowledge of inter-cue correlations but did not know how the different cues correlated with the criterion in consideration. Additionally, we systematically varied the cue-correlation matrix, and subsequently β_{EV1}, in order to adequately differentiate the predictions of this weighting scheme from those of alternate weighting schemes in our studies.

In our three studies, the multi-cue judgment task was presented as an advice integration task, with the cues in consideration corresponding to the judgments of four advisors. These cues were described as predicted stock prices in Studies 1 and 2 and restaurant ratings in Study 3. Correspondingly, the criterion variable was the true stock price in Studies 1 and 2 and the true restaurant quality in Study 3. The cue-criterion correlations were never revealed to the participants.

All studies consisted of three tasks. The first two tasks exposed the participants to the cues, so as to allow them to form mental representations of the cue space. The third task asked participants to predict the criterion value based on the cue values. Note that the participants' predictions made in Task 3 were in the same decision domain as in the first two tasks. In other words, there is no interdomain extrapolation in our experiment, and the cue spaces of the three tasks completely overlap with each other. In addition to being stated numerically, cue values in the three tasks were also shaded on the basis of their magnitude (darker shades for larger numbers). Participants were told explicitly that the cue values ranged from 0 to 100 and were all centered at 50. They were also told that some cues (advisors) might be more similar to each other, and that it was useful to pay attention to how closely different cues agreed with each other.

In Task 1, participants saw the four cues in 25 trials (Figure 2, upper left) displayed in four boxes. Each trial presented a set of cue values, and participants were asked to merely observe the cue values, without providing a response. In Task 2, participants continued to learn the cue values, this time with feedback. Particularly, in each trial, only three of the four cues were shown to participants (Figure 2, lower left). Participants had to guess the value of the fourth cue based on their learnt knowledge of the inter-cue relationships. After the participant's guess, the real cue value was revealed. There were 275 such trials, and the cue to be guessed was determined at random in each trial. To increase motivation, participants were provided with a summary of their performance accuracy after every 50 trials.

In Task 3, participants were shown all four cue values in each trial and were asked to make a guess regarding the value of the criterion (real stock price for Studies 1 and 2 and actual restaurant quality for Study 3; Figure 2, right). The true value of the criterion was not revealed after participants' guesses, so that participants stayed uninformed regarding the cue-criterion relationship. There were 100 such trials.

For all experiments, we used a single cue distribution for all three tasks for a given participant (though, of course, the cue distribution varied across different experiments and across conditions in a single experiment). Additionally, the order in which the trials were presented in each task was randomized across participants. Out of the three tasks, Task 3 was the most relevant to our research question, as it provided a direct test of how cue values were integrated to make a judgment of the criterion.

Note that although we are pursuing a key question in multi-cue judgment, that is, how inter-cue relationships influence cue integration in the absence of cue-criterion knowledge, our task also bears resemblance to the experimental paradigms used in the judge-advisor systems literature (see Bonaccio & Dalal, 2006; Yaniv, 2004, for reviews). To keep consistent with the multi-cue judgment literature, we refer to advisors' recommendations as *cues* in this paper. In Section 5, we consider the implications of our findings for the literature on judge-advisor systems, and the way in which existing research on judge-advisor systems can inform our understanding of cue integration in the absence of cue-criterion knowledge.

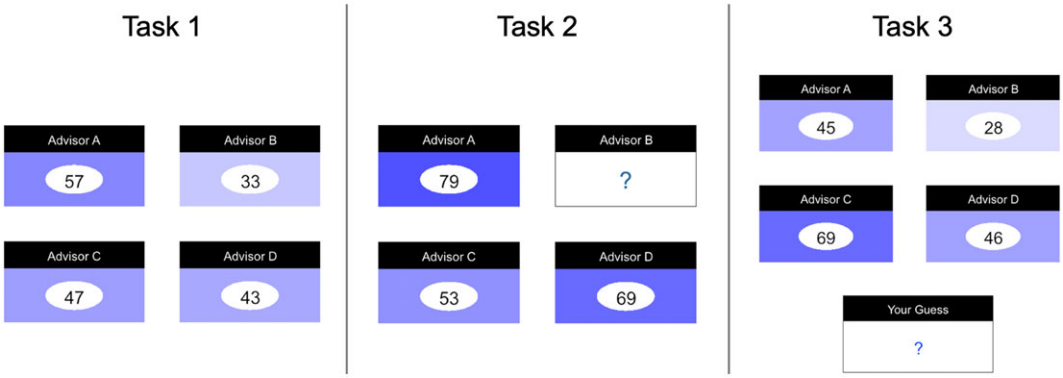

FIGURE 2 Stimuli display for Tasks 1–3. Task 1 (top left) involves the passive learning of cue correlations, whereas Task 2 (bottom left) involves the active learning of cue correlations. Task 3 (right) requires participants to aggregate cue values and predict the criterion variable. Our primary tests involve responses in Task 3 [Colour figure can be viewed at wileyonlinelibrary.com]

3.1 | Study 1

In Study 1, we wished to directly test whether participants used the EV1 weighting scheme introduced above. For this purpose, we selected a cue correlation matrix that led to a $\boldsymbol{\beta}_{EV1}$ weighting scheme, which placed a larger weight on one cue and smaller weights on the remaining cues. If participants optimally used inter-cue correlations to judge criterion values (in the absence of cue-criterion knowledge), they would place a larger weight on one of the cues relative to the others.

3.1.1 | Methods

Forty-four participants (37 females; Mean Age = 19.7, *SD* Age = 1.2), recruited from a university experimental participation pool, completed this study in a behavioral laboratory.

The study involved a hypothetical stock prediction task. The cue values were stock prices predicted by four advisors. For each cue, the values were normally distributed, with a mean of 50 and a standard deviation of 25. The cue correlation matrix of the advisors is shown in Figure 1a. As can be seen in this matrix, Cue 1 is highly correlated with all the three other cues, with a correlation coefficient of 0.6. The internal correlation among the remaining cues is very weak, with a correlation coefficient of 0.05. The eigenvector corresponding to the first (largest) eigenvalue of the cue correlation matrix is $\boldsymbol{\beta}_{EV1}$= [0.69, 0.42, 0.42, 0.42]. Using this weighting vector leads to an overweighting of Cue 1, and a relative underweighting of the remaining cues.

We used the above distributions to generate a single set of stimuli for all participants, for Tasks 1–3 (see the Supporting Information for summaries of the generated stimuli set). For each participant, the display position for each of the four cues was randomly chosen at the beginning of the study and stayed unchanged for the entire session. In other words, the specific advisor (advisor A, B, C, or D) associated with Cue 1 was randomized. Note that all participants were given the same set of trials in Tasks 1–3 (though the ordering of the trials was randomized within each task, for each participant).

3.1.2 | Modeling

The primary focus of this study was the weighting scheme used by participants for aggregating cues to predict criterion values in Task 3. For this purpose, we considered a number of candidate weighting schemes, including $\boldsymbol{\beta}_{EV1}$ = [0.69, 0.42, 0.42, 0.42] (corresponding to the first eigenvector of the cue-correlation matrix) and $\boldsymbol{\beta}_{EW}$ = [0.5, 0.5, 0.5, 0.5] (corresponding to the equal weighting rule). We also considered single cue heuristics. We tested four models that put all the weight on a single cue. These were referred to as $\boldsymbol{\beta}_{LEX1}$ = [1, 0, 0, 0], $\boldsymbol{\beta}_{LEX2}$ = [0, 1, 0, 0], $\boldsymbol{\beta}_{LEX3}$ = [0, 0, 1, 0], and $\boldsymbol{\beta}_{LEX4}$ = [0, 0, 0, 1]. In addition to $\boldsymbol{\beta}_{EV1}$, we also considered the linear weighting schemes corresponding to the remaining three eigenvectors of the cue correlation matrix. These are $\boldsymbol{\beta}_{EV2}$ = [0,0.77, − 0.63, − 0.14], $\boldsymbol{\beta}_{EV3}$ = [0, − 0.28, − 0.53,0.80], and $\boldsymbol{\beta}_{EV4}$ = [−0.72,0.40,0.40,0.40]. For this correlation matrix, we note that $\boldsymbol{\beta}_{EV2}$ and $\boldsymbol{\beta}_{EV3}$ are not uniquely determined; however, we include them as possible alternative weighting models orthogonal to $\boldsymbol{\beta}_{EV1}$. Therefore, we had, in total, nine improper linear weighting schemes to compare, with each weighting scheme defining a weighting vector for the four cues. Note that we cannot evaluate a "proper" linear model in which the cue weights are dependent on their validities, as the participants had no information about the cue-criterion relationships. However, in the Supporting Information, we discuss results regarding a flexible linear model, where there is no prespecified constraint on participants' weighting schemes.

Consider an improper weighting vector $\boldsymbol{\beta}$, and a vector of cue values $\boldsymbol{C}$ presented in a trial. In order to estimate each participant's linear representation, under the specified improper linear model constraints, we introduced three participant-level parameters, α_{0i}, α_{1i}, and σ_i, where i denotes participant number. We assumed that the participants' criterion predictions were drawn from a normal distribution ($-\alpha_{0i} + \alpha_{1i}\boldsymbol{\beta} \cdot \boldsymbol{C}$, σ_i). Here, α_{0i} captures a participant's general tendency to report higher (or lower) values, α_{1i} sets the scale of the weights (while $\boldsymbol{\beta} \cdot \boldsymbol{C}$ enforces the prespecified weighting relationships), and σ_i is the standard deviation of the normally distributed noise in the participant's responses. Based on the nine candidate weighting schemes we were interested in, we built nine (stochastic) models. In each of the nine models, α_{0i}, α_{1i}, and σ_i were fully flexible, and $\boldsymbol{\beta} \cdot \boldsymbol{C}$ was predetermined by the improper model, so each model had three parameters. Each model was fit separately to the 100 criterion predictions (in Task 3), for each of the 44 participants, using maximum likelihood estimation (see the Supporting Information for best-fit participant estimates). Our approach is tantamount to estimating the best-fitting linear model, for each participant, subject to the constraints on the weighting relationships determined by each of the nine fixed weighting schemes.

3.1.3 | Results

Learning cue correlations

We first examined participants' performance in Task 2, where they used three cue values for guessing the remaining cue value. In this task, participants who were able to learn the underlying cue relationships should place a higher weight on Cue 1 when guessing the values of Cues 2–4. We tested this using a linear regression analysis. For three subsets of trials, where participants predicted the cue values for Cues 2–4, we estimated three separate regression models with the guessed cue value as the dependent variable and the provided cue values as predictor variables. For example, in trials in which participants had to guess Cue 2, we regressed the guessed value onto the provided values of Cues 1, 3, and 4. The regression coefficients for the advisors obtained from the three regressions were averaged to generate a measurement of each participant's *dependence* on each cue. The average dependence on Cue 1, across participants, was 0.22 (*SD* = 0.19). In contrast, the average dependencies for Cues 2, 3, and 4 were 0.13 (*SD* = 0.14), 0.12 (*SD* = 0.15), and 0.13 (*SD* = 0.13). Additionally, among the 44 participants, 27 had the highest dependence on Cue 1, five had the highest dependence on Cue 2, six on Cue 3,

and six on Cue 4. Three paired participant-level Wilcoxon tests, which evaluated the difference between the dependence on Cue 1 relative to Cues 2–4, indicated that these differences were all statistically significant ($p < 0.05$ for all three tests). These analyses suggest that most participants learnt the special status of Cue 1 compared with the remaining cues. Note that using regression coefficients to compare cue importance may be problematic in some situations. An alternate approach is to perform a dominance analysis, which evaluates the contribution of each cue by examining the R^2 of all possible subset models (Azen & Budescu, 2003). The results of dominance analyses, applied to each of the four cues, confirmed the results of our regression analyses: the use of Cue 1 dominated Cues 2–4 for most participants, suggesting that they did learn the special status of Cue 1. Due to space constraints, these results are reported in the Supporting Information.

Note that participants who did not rely on Cue 1 more than the other cues in Task 2 also had lower performance in that task. We calculated the *relative dependence* on Cue 1 for each participant by taking the difference between the dependence on Cue 1 and the mean of the dependencies on the other cues. We measured performance using the mean squared guessing error across the trials. As expected, relative dependence was negatively correlated with the mean squared error (MSE; $corr = -0.55$, $t(42) = -4.25$, $p = 0.001$).

Cue weighting and aggregation

Our analysis of Task 2 revealed that participants were able to successfully learn inter-cue correlations (specifically, the special predictive status of Cue 1; see also the analysis of the relationship between Tasks 2 and 3 in the Supporting Information). How did they use this knowledge to weight cues when making a prediction on the criterion in Task 3? In order to test this, we first compared the fits of the nine candidate models at the group level. This analysis involved adding up the individual best-fitting log likelihood values for the 44 participants, for each model. Because all models have the same number of parameters, our model comparison is equivalent to model selection by the Akaike information criterion (Akaike, 1969) or the Bayes information criterion (Schwarz, 1978).

As shown in Table 1, we found that the best-fit model on the group level was the β_{EW} model, with an aggregate log likelihood of −15,712.32. The second-best model was β_{EV1}, which had an aggregate log likelihood of −15,715.19. This corresponds to a log likelihood difference of 2.87 relative to β_{EW}, which is typically considered to be a minor difference in model fit (Raftery, 1995). The third-best model was the β_{LEX1} model, with a log likelihood of −15,775.58. This corresponds to a log likelihood difference in excess of 60 relative to β_{EV1} and β_{EW}. This is typically considered to be a very large difference in relative model fit.

All the remaining models were much worse than the β_{LEX1} model (that is, had much lower log likelihood values). In line with these results, we found that on the group level the β_{EV1} model had similar MSE and mean absolute error (MAE) as the β_{EW} model. The model with the third smallest MSE and MAE was the β_{LEX1} model, whereas the other models had much larger MSE and MAE. These statistics are also shown in Table 1.

TABLE 1 Comparison of model fits for Study 1

Model	Aggregate LL	Median LL	Median MSE	Median MAE	# Best
EV1	−15,715.19	−351.38	73.40	6.91	10
EV2	−18,504.75	−424.01	348.37	15.49	1
EV3	−18,508.27	−423.86	346.98	15.33	1
EV4	−18,506.19	−424.05	347.22	15.58	0
EW	−15,712.32	−351.59	73.24	6.88	23
LEX1	−15,775.58	−351.39	76.68	6.97	8
LEX2	−17,884.02	−408.81	250.66	12.32	0
LEX3	−17,918.43	−410.32	254.38	13.01	0
LEX4	−17,856.75	−408.70	242.65	12.59	1

Note. Here, aggregate LL corresponds to the sum of the best-fit log likelihood values across participants. Likewise, median LL corresponds to the median best-fit log likelihood value across participants. Median MSE corresponds to the median of mean squared error calculated from the best-fit model, whereas median MAE corresponds to the median of mean absolute error across participants; "# Best" indicates the number of participants best fit according to log likelihood, by a given model.

Using only aggregate analysis to infer individual behavioral patterns can be problematic (e.g., Estes, 1956). Therefore, we also compared model fits at the individual level. We compared the log likelihood of each model for every participant separately. Among the 44 participants, 10 participants' predictions were best described by β_{EV1}, 23 by β_{EW}, eight by β_{LEX1}, one by β_{EV2}, one by β_{EV3}, and one by β_{LEX4} (see Table 1 for more details). Next, we tested the relative model fits for pairs of models. For this purpose, we used paired Wilcoxon signed-rank tests to examine whether the differences in model fits across participants follow a symmetrical distribution centered at 0. There was no significant difference between log likelihood values of the β_{EV1} model ($Median = -351.38$) and the β_{EW} model ($Median = -351.59$; $z = 1.23$, $p = 0.225$). However, fits for all of the remaining models were significantly worse than those for the β_{EV1} model and β_{EW} model ($p < 0.001$ for all pairwise comparisons with β_{EV1} and β_{EW}).

The above results suggest that although the overall fits for β_{EV1} and β_{EW} were nearly identical, only 10 (out of 44) participants were best described by β_{EV1}. But note, of course, that β_{LEX1}, the third best model, is a more extreme form of β_{EV1}, which places all the weight on the exceptional cue (Cue 1). Thus, participants best fit by β_{LEX1} can actually be seen as utilizing β_{EV1} in a more extreme manner than would be prescribed to minimize maximal risk (the other three lexicographic rules each only best described the behavior of one participant, indicating that β_{EV1} can predict which single cue participants tend to overweight). Indeed, when directly comparing β_{EV1} and β_{EW} with each other, we found that 19 participants were better described by β_{EV1} (these 19 included the eight participants previously best described by β_{LEX1}), whereas 25 were better described by β_{EW}.

We also calculated the relative likelihood of the β_{EV1} model compared with the β_{EW} model conditional on the observed data, assuming the prior probabilities of β_{EV1} and β_{EW} model were both 0.5. The

method was taken from Wagenmakers and Farrell (2004): Using the log likelihood of models fit to each participant's data, we can calculate how much evidence the data provided for each model in the form of a likelihood. As can be seen in Figure 3a, the relative likelihood of β_{EV1} is noticeably larger than β_{EW} for a substantial subgroup of participants. Similar pairwise comparisons between β_{EV1} and the remaining weighting schemes are shown in Figure 3b–h. These figures again indicate that β_{EV1} fit the data much better than the remaining weighting schemes for a large group of participants.

3.1.4 | Discussion

Study 1 showed that most participants were able to learn inter-cue correlations in our judgment task, and that a substantial subgroup of participants used this knowledge to integrate cue values. Although these participants had absolutely no information regarding the validities of any of the cues, they did not appear to simply average the cue values. Instead, many of these participants tended to overweight Cue 1 as predicted by β_{EV1}. In fact, a number of participants were actually best fit by β_{LEX1} indicating that these participants overweighed Cue 1 even more than β_{EV1} recommended. On the aggregate level, β_{EV1} and β_{EW} fit the data about equally well and were statistically indistinguishable in terms of relative model fit.

3.2 | Study 2

Although Study 1 showed that a number of participants appeared to be behaving in accordance with the predictions of β_{EV1}, it involved an important limitation. Notably, the cue correlation matrix was not varied across participants. Thus, it may be the case that participants' apparent reliance on β_{EV1} was not due to the (optimal) decision to weight cues based on inter-cue correlations, but rather due to another, incidental or exogenous behavioral tendency.

The goal of Study 2 was to control for this potential confound by systematically varying the cue correlation matrix across participants. For this purpose, it presented half of the participants with the asymmetric cue correlation matrix in Study 1, for which β_{EV1} accords a special status to Cue 1, and presented the other half of the participants with a symmetric cue correlation matrix, for which β_{EV1} weighs all cues equally.

3.2.1 | Methods

Sixty-four participants (35 females; Mean Age = 19.9, *SD* Age = 1.1), recruited from a university experimental participation pool, completed this study in a behavioral laboratory.

All aspects of the study design were kept identical to Study 1, except that the cue correlation matrix varied between a treatment condition and a control condition. Participants were randomly assigned to one of these two conditions at the start of the study. For the treatment condition, the cue correlation matrix was identical to that in Study 1 (Figure 1a), generating an optimal weighting scheme with $\beta_{EV1}^{Treat} = [0.69, 0.42, 0.42, 0.42]$ (here, we use the superscript to distinguish the treatment vs. control condition). For the control condition, the cue correlation matrix kept the correlation between all the cues constant at 0.4 (Figure 1b). Therefore, the weighting vectors predicted by the optimal weighting scheme and the equal weights rule were both $\beta_{EV1}^{Cont} = \beta_{EW} = [0.5, 0.5, 0.5, 0.5]$. Due to different inter-cue relationships across conditions, β_{EV1}^{Treat} should provide a better account of behavior in the treatment condition compared with the control condition. Likewise, $\beta_{EV1}^{Cont} = \beta_{EW}$ should provide a better

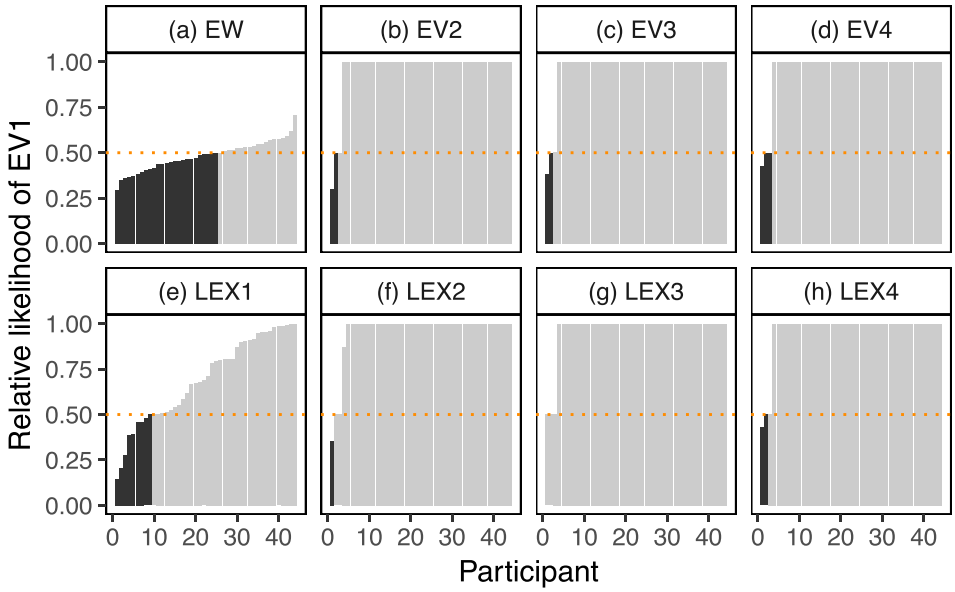

FIGURE 3 (a–h) The relative likelihood of the β_{EV1} model in Study 1, compared with the other eight models, for each of the participants. Here, lighter bars indicate participants that are overall better fit by β_{EV1}, and darker bars indicate participants that are overall better fit by the alternate model in consideration [Colour figure can be viewed at wileyonlinelibrary.com]

account of behavior in the control condition compared with the treatment condition (even if a large subgroup of participants in the treatment condition do place equal weight on all cues).

Again, all the weighting schemes used in this study were embedded in a normally distributed error model as described above and fit using maximum likelihood estimation applied at the participant level.

3.2.2 | Results

Learning cue correlations

Thirty-one participants were assigned to the treatment condition, and 33 participants were assigned to the control condition. As in Study 1, we first looked at participant learning in Task 2. In the treatment condition, the average dependence on Cue 1 was 0.20 (*SD* = 0.16), and the average dependencies on Cues 2–4 were 0.13 (*SD* = 0.15), 0.14 (*SD* = 0.13), and 0.10 (*SD* = 0.17), respectively. Additionally, when ranking the dependencies on the four cues for each participant, we found that among the 31 participants, 14 had the highest overall dependence on Cue 1, five had highest overall dependence on Cue 2, four on Cue 3, and eight on Cue 4. We again used paired Wilcoxon tests to evaluate the difference between the dependence on Cue 1 and the dependencies on Cues 2–4. Although most participants did place a higher relative weight on Cue 1, not all of these differences reached statistical significance, presumably due to lack of power ($p = 0.14$; $p = 0.11$; $p = 0.04$, for comparisons between Cue 1 and Cues 2–4). Finally, as in Study 1, there was a negative correlation between the relative dependence on Cue 1 and performance error (corr = − 0.41, t(29) = − 2.41, p = 0.023).

In the control condition, the overall dependence on the four cues was 0.22 (*SD* = 0.06), 0.22 (*SD* = 0.09), 0.25 (*SD* = 0.10), and 0.24 (*SD* = 0.10), respectively. Among the 33 participants, five, eight, 11, and nine participants placed the highest weights on Cues 1–4, respectively. Note that here there is no special status of any given cue, and all cues are distributionally identical (thus, our labeling of a given cue as 1, 2, 3, or 4 is arbitrarily based on the order in which the cue values were labeled in the experimental software program we used to generate stimuli). We repeated these regression-based cue dependence tests with dominance analysis (Azen & Budescu, 2003), which revealed similar results. We present the results of the dominance analysis in the Supporting Information.

Cue weighting and aggregation

The successful learning of inter-cue correlations in Task 2 of the treatment condition laid the basis for our analysis of Task 3, in which participants were required to aggregate cues to predict criterion values.

In the treatment condition, we replicated the results found in Study 1. On the aggregate level, the best-fit model was $\boldsymbol{\beta_{EW}}$ (Table 2). The $\boldsymbol{\beta}_{\boldsymbol{EV1}}^{\text{Treat}}$ model's fits was again closest to the best-fit model, with a log likelihood difference of 6.12. The third-best model was still $\boldsymbol{\beta_{LEX1}}$, with a log likelihood difference of more than 55 relative to $\boldsymbol{\beta_{EW}}$ and $\boldsymbol{\beta}_{\boldsymbol{EV1}}^{\text{Treat}}$. All other models tested were much worse than the $\boldsymbol{\beta_{LEX1}}$ model by this measure. Similar results were obtained from analyses of MSE and MAE, where $\boldsymbol{\beta}_{\boldsymbol{EV1}}^{\text{Treat}}$, $\boldsymbol{\beta_{EW}}$, and $\boldsymbol{\beta_{LEX1}}$ had smaller prediction errors than the remaining models.

We also compared the log likelihood values of the fits on the individual level. Although the log likelihood values of the $\boldsymbol{\beta}_{\boldsymbol{EV1}}^{\text{Treat}}$ model were significantly smaller than those of the $\boldsymbol{\beta_{EW}}$ model ($Z = -2.06$, $p = 0.040$), the differences were minor ($Median_{EV1} = -350.56$, $Median_{EW} = -350.17$). Additionally, both the $\boldsymbol{\beta}_{\boldsymbol{EV1}}^{\text{Treat}}$ model and $\boldsymbol{\beta_{EW}}$ models had higher log likelihood values than all other models ($p < 0.001$ for all pairwise comparisons with $\boldsymbol{\beta_{EW}}$ and $\boldsymbol{\beta}_{\boldsymbol{EV1}}^{\text{Treat}}$). Out of 31 participants in the treatment condition, six were best described by $\boldsymbol{\beta}_{\boldsymbol{EV1}}^{\text{Treat}}$, 18 by $\boldsymbol{\beta_{EW}}$, six by $\boldsymbol{\beta_{LEX1}}$, and one by $\boldsymbol{\beta_{EV2}}$. As in Study 1, some participants were best described by $\boldsymbol{\beta_{LEX1}}$, indicating that they overweighed Cue 1 more than recommended by $\boldsymbol{\beta}_{\boldsymbol{EV1}}^{\text{Treat}}$.

When comparing only $\boldsymbol{\beta}_{\boldsymbol{EV1}}^{\text{Treat}}$ and $\boldsymbol{\beta_{EW}}$, 13 out of 31 participants were better described by the $\boldsymbol{\beta}_{\boldsymbol{EV1}}^{\text{Treat}}$ model. This difference is illustrated in Figure 4a. Here, the relative likelihood of $\boldsymbol{\beta_{EV1}}$ is larger than $\boldsymbol{\beta_{EW}}$ for a substantial subgroup of participants. Pairwise comparisons between $\boldsymbol{\beta_{EV1}}$ and the remaining weighting schemes are shown in Figure 4b–h. These figures again indicate that $\boldsymbol{\beta_{EV1}}$ fit the data much better than the remaining weighting schemes.

In the control condition, the cue correlation matrix was balanced and $\boldsymbol{\beta}_{\boldsymbol{EV1}}^{\text{Cont}}$ was identical to $\boldsymbol{\beta_{EW}}$. The linear weighting schemes corresponding to the other three eigenvectors of the cue correlation matrix were $\boldsymbol{\beta}_{\boldsymbol{EV2}}^{\text{Cont}} = [0, -0.26, -0.54, 0.80]$, $\boldsymbol{\beta}_{\boldsymbol{EV3}}^{\text{Cont}} = [0.46, 0.51, -0.67, -0.29]$, and $\boldsymbol{\beta}_{\boldsymbol{EV4}}^{\text{Cont}} = [0.74, -0.65, 0.08, -0.16]$. These three eigenvector weighting schemes are not uniquely determined by this correlation matrix, but, similar to before, we include them as competitor weighting schemes orthogonal to $\boldsymbol{\beta}_{\boldsymbol{EV1}}^{\text{Cont}}$. Unsurprisingly, all 33 participants were

TABLE 2 Comparison of model fits for Study 2's treatment condition

Model	Aggregate LL	Median LL	Median MSE	Median MAE	# Best
$EV1_{Treat}$	−11269.97	−350.56	101.44	7.46	6
EV2	−13123.40	−424.88	389.37	16.29	1
EV3	−13131.69	−425.08	389.22	16.34	0
EV4	−13129.12	−424.85	386.54	16.28	0
EW	−11263.85	−350.17	101.35	7.48	18
LEX1	−11325.32	−353.77	103.90	7.30	6
LEX2	−12671.19	−407.65	265.18	12.62	0
LEX3	−12766.85	−412.35	298.54	13.78	0
LEX4	−12693.89	−411.66	275.24	13.34	0

Note. Here, aggregate LL corresponds to the sum of the best-fit log likelihood values across participants. Likewise, median LL corresponds to the median best-fit log likelihood value across participants. Median MSE corresponds to the median of mean squared error calculated from the best-fit model, whereas median MAE corresponds to the median of mean absolute error across participants; "# Best" indicates the number of participants best fit according to log likelihood, by a given model.

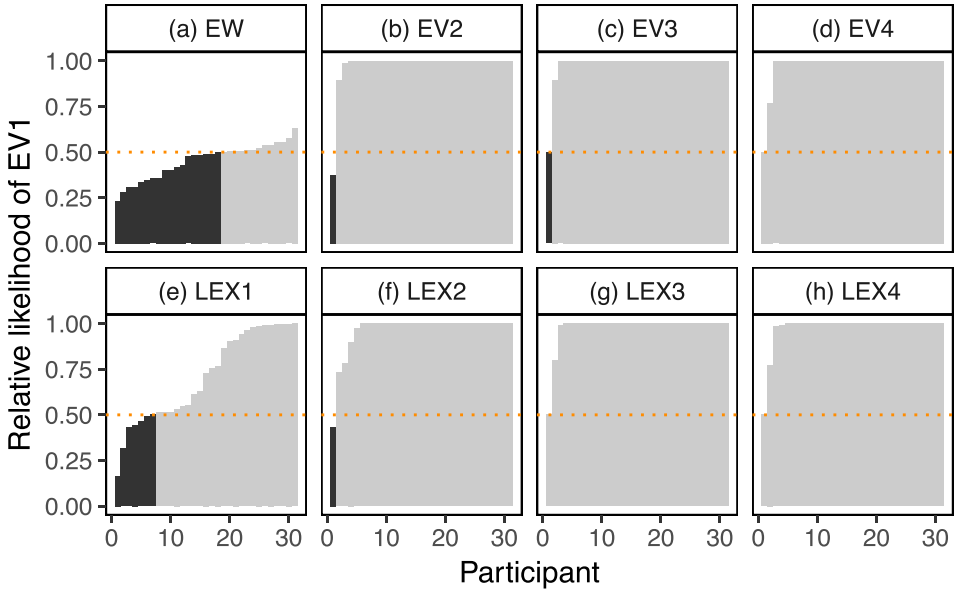

FIGURE 4 (a–h) The relative likelihood of β_{EV1} model in the treatment condition of Study 2, compared with the other eight models, for each of the participants. Here, lighter bars indicate participants that are overall better fit by β_{EV1}, and darker bars indicate participants that are overall better fit by the alternate model in consideration [Colour figure can be viewed at wileyonlinelibrary.com]

better described by $\beta_{EV1}^{Cont} = \beta_{EW}$ compared with the other models (Table 3; Figure 5a–g).

Lastly, we examined the predictions of β_{EV1}^{Treat} on the data from the control condition. This was done to test whether the descriptive power of β_{EV1}^{Treat} relative to β_{EW} does in fact diminish when β_{EV1}^{Treat} no longer describes the cue correlation structure. This involved fitting a tenth model in the control condition, with weights given by $\beta_{EV1}^{Treat} = [0.69, 0.42, 0.42, 0.42]$. Note that due to cue symmetry, it is not clear which cue should be assigned the higher weight. Thus, we assumed that participants randomly chose a cue to be "special" and assigned the extra weight to that cue in order to generate their responses. We simulated this process 10,000 times and computed the log likelihood of the tenth model as an average of the models fitted in the 10,000 simulations. Unlike the treatment condition, this model did not fit better than the $\beta_{EW} = \beta_{EV1}^{Cont}$ model for any participants. A paired Wilcoxon test indicated that the log likelihood of the β_{EV1}^{Treat} model on the control-condition data (*Median* = −335.76) was significantly lower than that of β_{EW} (*Median* = −339.20; $z = -5.01$, $p < 0.001$). The relative likelihood of the β_{EV1}^{Treat} model compared with the β_{EW} model is presented in Figure 5h. Similar results were obtained from the analyses of MSE and MAE. As can be seen in Table 3, the β_{EV1}^{Treat} model had larger MSE and MAE than the β_{EW} model, or equivalently, the β_{EV1}^{Cont} model.

TABLE 3 Comparison of model fits for Study 2's control condition

Model	Aggregate LL	Median LL	Median MSE	Median MAE	# Best
$EV1_{Control}$	−11092.29	−335.76	56.09	5.80	33
EV2	−13593.19	−415.22	311.25	14.29	0
EV3	−13593.28	−415.22	311.09	14.30	0
EV4	−13585.70	−414.84	310.89	14.13	0
EW	−11092.29	−335.76	56.09	5.80	33
LEX1	−12713.27	−387.19	168.94	10.32	0
LEX2	−12830.93	−391.31	176.82	10.77	0
LEX3	−12610.18	−382.23	148.81	9.76	0
LEX4	−12787.58	−390.59	177.35	10.57	0
$EV1_{Treat}$	−11211.90	−339.20	106.83	8.40	0

Note. Here, aggregate LL corresponds to the sum of the best-fit log likelihood values across participants. Likewise, median LL corresponds to the median best-fit log likelihood value across participants. Median MSE corresponds to the median of mean squared error calculated from the best-fit model, whereas median MAE corresponds to the median of mean absolute error across participants; "# Best" indicates the number of participants best fit according to log likelihood, by a given model.

3.2.3 | Discussion

Study 2 tested whether the weighting schemes used by participants varied as a function of the cue correlations in the judgment task. In the treatment condition of Study 2, the cue correlation matrix was identical to that used in Study 1. Here, a substantial group of participants appeared to use the recommended β_{EV1} weighting scheme, which overweighted Cue 1. This replicated the results of Study 1. None of the participants, however, used this weighting scheme in the control condition, where the cue correlation matrix was symmetric, and β_{EV1}weighted all cues equally.

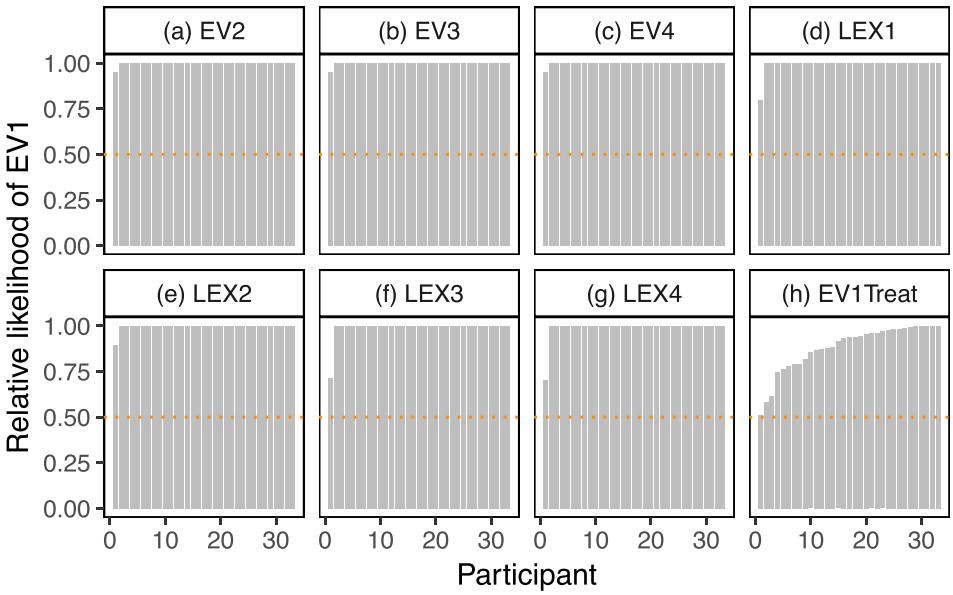

FIGURE 5 (a–h) The relative likelihood of $\beta_{EV1}^{Control}$ model (which is equivalent to the β_{EW} model) in the control condition of Study 2, compared with the other eight models, for each of the participants. Here, lighter bars indicate participants that are overall better fit by $\beta_{EV1}^{Control} = \beta_{EW}$ [Colour figure can be viewed at wileyonlinelibrary.com]

3.3 | Study 3

Studies 1 and 2 suggest that some participants were able to optimally use inter-cue relationships when criterion values were unknown. In Study 3, we wished to test the boundaries of this effect, by considering a setting with a highly complex cue correlation matrix. Study 3 also used judgments of restaurant quality rather than stock performance, as stimuli, to ensure that the results did not depend on the precise judgment domain.

3.3.1 | Methods

Forty-six participants (34 females; Mean Age = 19.3, *SD* Age = 1.0) recruited from a university experimental participation pool, completed this study in a behavioral laboratory.

The study was framed as involving judgments of restaurant quality. Here, the cue values were restaurant scores rated by four reviewers, and the criterion corresponded to the true restaurant quality. The cue correlation matrix used was the one displayed in Figure 1c. Here, Cue 1 is highly correlated with Cue 2, Cue 3 is moderately correlated with Cue 4, and Cues 1 and 2 are weakly correlated with Cues 3 and 4. With this cue correlation structure, β_{EV1} predicts a weighting vector of [0.65, 0.65, 0.27, 0.27], that is, an overweighting of Cues 1 and 2, relative to 3 and 4. The weighting schemes corresponding to the other eigenvectors are β_{EV2} = [−0.27, − 0.27,0.65,0.65], β_{EV3} = [0, 0, .071, − 0.71], and β_{EV4} = [0.71, − 0.71,0, 0]. These three eigenvector schemes are uniquely determined. Other aspects of the study design and model fitting were identical to Study 1.

3.3.2 | Results

Learning cue correlations

As in Study 1, we first tested whether participants were able to learn the cue correlation matrix. Again note that Cues 1 and 2 are strongly correlated and Cues 3 and 4 are moderately correlated. Thus, when predicting the value of Cue 1 (Cue 2), participants should rely more on Cue 2 (Cue 1) compared with Cues 3 and 4. Conversely, when predicting the value of Cue 3 (Cue 4), participants should rely more on Cue 4 (Cue 3) compared with Cues 1 and 2.

We did find evidence for this type of learning. In trials in which participants were required to predict Cues 1 and 2, the average dependencies for Cues 1–4 were 0.69 (*SD* = 0.23), 0.66 (*SD* = 0.24), 0.06 (*SD* = 0.01), and 0.09 (*SD* = 0.10), respectively. Additionally, the number of participants who put the highest weights on Cues 1–4 were 25, 19, 0, and 2, respectively. Paired Wilcoxon tests revealed that the dependencies for Cues 1 and 2 were higher than those for Cues 3 and 4 (z =10.91, $p < 0.001$). In contrast, in trials in which participants were required to predict Cues 3 and 4, the average dependencies for Cues 1–4 were 0.11 (*SD* = 0.12), 0.07 (*SD* = 0.11), 0.57 (*SD* = 0.22), and 0.58 (*SD* = 0.23), respectively. Additionally, the number of participants who put the highest weights on Cues 1–4 were 4, 0, 18, and 24, respectively. In this setting, paired Wilcoxon tests revealed that the dependencies for Cues 3 and 4 were higher than those for Cues 1 and 2 (z =10.53, $p < 0.001$). Finally, as in Study 1 and the treatment condition of Study 2, we found that there was a significant negative correlation between people's dependencies on the cues and their guessing error ($corr$ = − 0.72 $t(44)$ = − 6.88, $p < 0.001$). The Supporting Information present the results of dominance analyses on these data (Azen & Budescu, 2003). These results confirm the findings outlined here.

Cue weighting and aggregation

We investigated the linear weighting schemes used by participants in Task 3. The nine candidate weighting schemes and the model fitting procedures were the same as in Studies 1 and 2 (except that β_{EV} was obtained from the cue correlation matrix in Figure 1c). On the aggregate level, the best-fit model was the β_{EW} model (Table 4). The second-best model was again the β_{EV1} model. The log likelihood difference between the β_{EW} model and the β_{EV1} model in this study was 421.7, which is much larger than the equivalent difference in Study 1 and the treatment condition of Study 2. However, as in these experiments, the β_{EW} model and the β_{EV1} model fit much better than the other seven models. Similar results were obtained from the analyses of MSE and MAE.

TABLE 4 Comparison of model fits for Study 3

Model	Aggregate LL	Median LL	Median MSE	Median MAE	# Best
EV1	−16480.03	−353.15	86.21	6.92	7
EV2	−18887.24	−412.07	302.28	13.81	0
EV3	−19026.78	−415.25	298.96	13.84	0
EV4	−19035.65	−415.29	302.28	13.81	0
EW	−16058.33	−350.82	76.66	6.45	38
LEX1	−17694.89	−383.14	158.33	10.09	0
LEX2	−17539.22	−379.23	142.68	9.47	1
LEX3	−18490.79	−403.08	234.83	11.97	0
LEX4	−18316.33	−400.57	213.93	11.56	0

Note. Here, aggregate LL corresponds to the sum of the best-fit log likelihood values across participants. Likewise, median LL corresponds to the median best-fit log likelihood value across participants. Median MSE corresponds to the median of mean squared error calculated from the best-fit model, whereas median MAE corresponds to the median of mean absolute error across participants; "# Best" indicates the number of participants best fit according to log likelihood, by a given model.

On the individual level, β_{EW} (*Median* = −350.82) was significantly better than β_{EV1} (*Median* = −353.15) according to a paired Wilcoxon test performed on participant-level log likelihood values ($z = 3.98$, $p < 0.001$). Additionally, β_{EW} and β_{EV1} fit better than all the other candidate models ($p < 0.001$). Out of the 46 participants, seven were best fit by β_{EV1} and 38 were best fit by β_{EW}. The remaining participant was best fit by β_{LEX2}. When comparing only β_{EV1} and β_{EW}, we found that eight participants were better fit by β_{EV1} compared with β_{EW}. This finding is illustrated in Figure 6a. For the eight participants who were better fit by the β_{EV1} model, the relative likelihood of the β_{EV1} model is above 0.97, and the relative likelihood of β_{EW} is less than 0.03 (indicating strong evidence for the few participants that were actually better fit by β_{EV1}). The comparisons between β_{EV1} and the other models are presented in Figure 6b–h. These figures show that β_{EV1} fit better than all other models for almost all participants.

3.3.3 | Discussion

As in previous studies, we found that there was a subgroup of participants that overweighed some cues (as suggested by β_{EV1}), rather than simply averaging all the available cues (as suggested by β_{EW}). However, the size of the subgroup was much smaller than in the previous studies. This is likely a result of the complexity of the cue-correlation matrix. It seems that participants are less likely to use inter-cue relationships optimally in settings in which these relationships involve multiple pairs of differentially correlated cues.

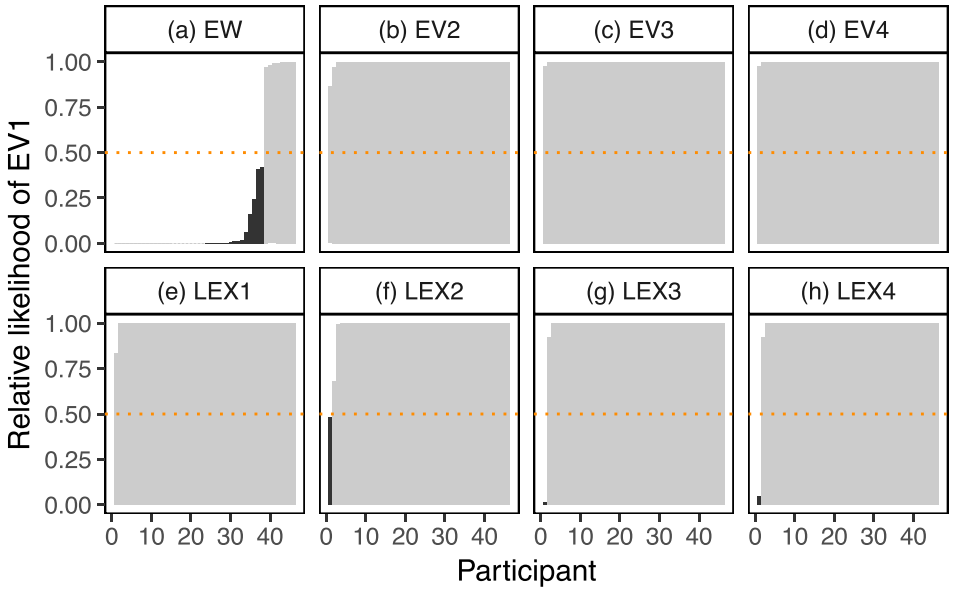

FIGURE 6 (a–h) The relative likelihood of the β_{EV1} model in Study 3, compared with the other eight models, for each of the participants. Here, lighter bars indicate participants that are overall better fit by β_{EV1}, and darker bars indicate participants that are overall better fit by the alternate model in consideration [Colour figure can be viewed at wileyonlinelibrary.com]

4 | MODEL RECOVERY STUDY

The above three studies suggest that some participants appear to be optimally using inter-cue relationships when predicting criterion values to minimize maximal risk. The tests in these studies involve model fitting and comparison for different types of weighting schemes using maximum log likelihood, and most of the claims in this paper rely on the finding that the optimal weighting scheme has a higher best-fit log likelihood for some participants relative to other improper linear models, such as the equal weights rule.

There is one possibility that the above tests have not taken into account. It is possible that β_{EV1} and β_{EW} models actually make very similar predictions for the types of stimuli examined in this paper, and that it is thus difficult to appropriately discriminate between the two models using model fitting to participant data. Consider, for example, the cue correlation matrix used in Study 1 and in the treatment condition of Study 2 (Figure 1a). Here, one of the cues correlates highly with the other cues, whereas all of the other cues are weakly correlated with each other. This correlation structure implies that overweighting the special cue (as recommended by the β_{EV1} weighting scheme) can actually yield predictions that are close to the average of all the cue values. For this reason, the differences between the predictions of the β_{EV1} model and the β_{EW} model could be relatively small, implying that our statistical tests could have misclassified participants using β_{EV1} as those using β_{EW} and vice versa.

To test for the possibility of this misclassification, we conducted a model recovery study using the medians of the individual level parameters α_{0i}, α_{1i}, and σ_i for participants best fit by the β_{EV1} model and participants best fit by the β_{EW} model in the treatment condition of Study 2. We used these median parameter values to generate hypothetical β_{EV1} and β_{EW} decision makers, whose judgments we simulated on the 100 trials in Task 3 of the treatment condition of Study 2. Each hypothetical decision maker was simulated 10,000 times. For each simulation, we fit the simulated data with both the β_{EV1} model and the β_{EW} model and determined which model performed better. Overall, the two models were recoverable, that is, data simulated with the β_{EV1} weighting scheme was more likely to be better fit by the β_{EV1} model compared with the β_{EW} model, and vice versa. However, the misclassification rates were also quite high. On average, 44% of β_{EV1} decision maker simulations were misclassified as β_{EW}. Likewise, 44% of β_{EW} decision maker simulations were misclassified as β_{EV1}. The relative likelihoods for the two models for the two sets of simulations are shown in Figure 7a. These results suggest that the task used in Study 1 and the treatment condition of Study 2 may not be ideal for distinguishing the predictions of β_{EV1} and β_{EW}. It could be the case that some of the β_{EV1} decision makers in our study were incorrectly determined to be β_{EW} decision makers and that some of the β_{EW} decision makers in our study were incorrectly determined to be β_{EV1} decision makers.

We also performed such model recovery studies for the cue correlation matrices used in the control condition of Studies 2 and 3. Again, for this purpose, we simulated hypothetical β_{EV1} and β_{EW} decision makers 10,000 times using median parameter values recovered from the control condition of Studies 2 and 3, respectively. Note that the β_{EV1} model simulated and fit in the control condition of Study 2 is the β_{EV1} from the treatment condition. As shown in Figure 7b,c, the misclassification rates were relatively low for both models in the control condition of Study 2 (<17%) and very low for both models in Study 3 (<1%). Thus, accurate model recovery is less likely to be a problem in Study 3 and the control condition of Study 2.

5 | DISCUSSION

In this paper, we investigated how decision makers integrate cues in a decision context where cue-criterion relationships are unknown. The optimal improper linear model in this setting involves weights determined by the eigenvector, β_{EV1}, corresponding to the largest

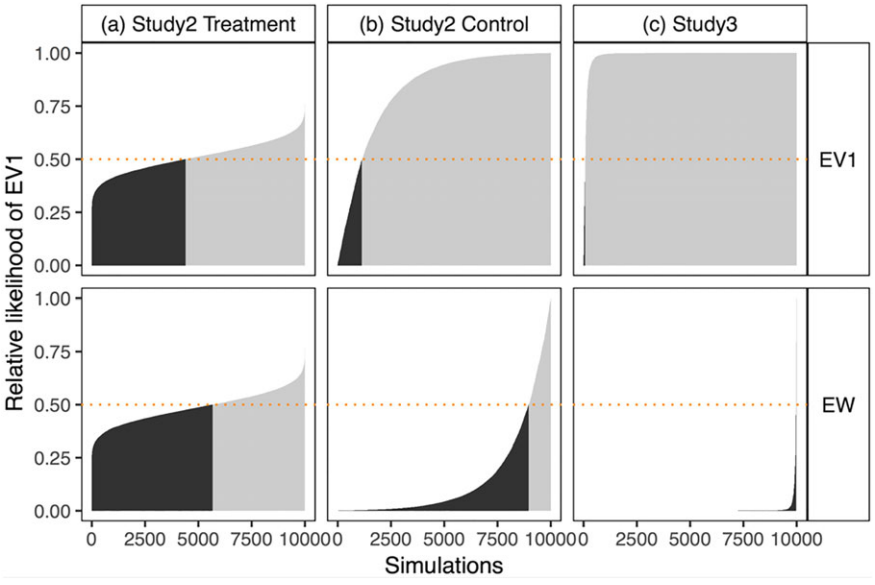

FIGURE 7 (a–c) Model recovery: Relative likelihoods for the β_{EV1} model calculated from the simulated data (upper row: using the β_{EV1} model as the data generating model; lower row: using the β_{EW} model as the data generating model. Note that the β_{EV1} model simulated and fit in the control condition of Study 2 is the β_{EV1} from the treatment condition). In the treatment condition of Study 2, the misclassification rate was high; in the control condition of Study 2 and Study 3, the misclassification rate was low [Colour figure can be viewed at wileyonlinelibrary.com]

eigenvalue of the cue correlation matrix (Davis-Stober et al., 2010a, 2010b). Low-dimensional representations of the cue space, learnt by some common models of semantic memory (Kruskal & Wish, 1978; Landauer & Dumais, 1997; Saxe et al., 2013), can also produce this type of weighting scheme.

We conducted three studies to examine whether decision makers utilize knowledge of inter-cue relationships to predict criterion values. Our studies consisted of three tasks: In the first two tasks, participants learnt inter-cue correlations. In the third task, they used their knowledge of the cue correlations to predict criterion values. Participants were never shown true criterion values and thus had no way of identifying cue-criterion validities.

We use model comparisons to test the use of cues in predicting criterion values. Our results from Task 3 of our studies suggest that some participants did use cue correlation information in a manner suggested by the optimal linear model that minimizes maximum risk. On both the aggregate and the individual level, $\boldsymbol{\beta_{EV1}}$ fit better than all other improper linear models tested in this paper, except for the equal weights model (with weights $\boldsymbol{\beta_{EW}}$). On the aggregate level, the log likelihoods for the $\boldsymbol{\beta_{EV1}}$ and $\boldsymbol{\beta_{EW}}$ showed no meaningful differences in Study 1, very minor differences in the treatment condition of Study 2, and somewhat larger differences in Study 3. As for individual level fits, there existed a substantial group of participants for whom $\boldsymbol{\beta_{EV1}}$ fit better than $\boldsymbol{\beta_{EW}}$. The size of this group ranged from 17% of the participant pool in Study 3, to 43% in Study 1 and 42% in the treatment condition of Study 2. Moreover, a comparison of the control and the treatment conditions of Study 2 showed that experimental manipulations that varied the cue correlation matrix influenced relative model fits.

$\boldsymbol{\beta_{EV1}}$ was also able to predict when and how participants used lexicographic weights. When $\boldsymbol{\beta_{EV1}}$ prescribed equal weights (control condition of Study 2) or overweighting two cues (Study 3), there were almost no participants who were best described by lexicographic weighting schemes. In contrast, in Study 1 and the treatment condition of Study 2, $\boldsymbol{\beta_{EV1}}$ overweighed a single cue. In these conditions, a substantial group of participants (18% in Study 1 and 19% in the treatment condition of Study 2) behaved in a way better described by a lexicographic rule that placed all of the weight on this cue (note that lexicographic rules that prioritize other cues all fit very poorly).

Our finding that some participants were able to appropriately utilize inter-cue correlations to predict criterion values is congruent with prior work that shows that cue correlations can influence learning, judgment, and confidence in multi-cue judgment tasks (Armelius & Armelius, 1976; Broomell & Budescu, 2009; Budescu et al., 2003; Budescu & Rantilla, 2000; Sniezek & Buckley, 1995). Our finding is also consistent with prior research that suggests that decision makers use heuristic decision strategies (such as improper linear models) when such strategies are adaptively rational (Anderson, 1990; Gigerenzer & Gaissmaier, 2011; Griffiths & Tenenbaum, 2016; Oaksford & Chater, 2007).

Note that we used an advice integration task to examine a question in multi-cue judgment. Although essentially the two types of tasks both involve combining multiple pieces of information, people's prior beliefs about inter-cue correlations can be different between the two tasks. Future research should examine how such prior beliefs interact with additional knowledge on inter-cue correlations acquired from an experimental setting, when people use them to predict criterion values. Moreover, although we avoid giving participants cue validity information for experimental control, it would be interesting to test how inter-cue correlations and cue-validity information combine to guide judgment, and whether our results persist even with some knowledge about cue validity.

Although the goal of this paper was to examine cue-integration in an abstract multi-cue judgment task, the social nature of our stimuli implies that our findings relate to some of the core questions of interest in the judge-advisor systems literature, which studies how people solicit, weigh and aggregate the advice of others (Bonaccio & Dalal, 2006; Yaniv, 2004). Previous work in this area has shown that judges are sensitive not only to the features of a single advisor but also to advisor interrelationships (see, e.g., Broomell & Budescu, 2009; Budescu et al., 2003; Budescu & Rantilla, 2000; Sniezek & Buckley, 1995). In the current paper, we selectively test the influence of advisor interrelationships on advice integration, by withholding information regarding advisor accuracy from participants. Our results indicate that judges' weighting strategies might depend on their prior knowledge of advisor similarities.

There are also some results in the judge-advisor systems literature that could be better understood through the lens of our multi-cue judgment model. For example, prior work has found that judges discount extreme advice in an advice integration task (Harries, Yaniv, & Harvey, 2004; Yaniv, 1997). It could be the case that judges are able to discover correlation structures underlying different advice sources and optimally apply this knowledge to discount the weakly correlated advisor. That said, there are also some important differences between prior findings regarding the discounting of extreme advisors and our finding regarding the influence of inter-cue correlations on cue integration. Particularly, Harries et al. (2004) define extreme advice on a trial-to-trial basis, so that advisors deviating from the mean of the remaining advisors in a single trial are considered outliers. In contrast, in our task, (dis)similarity among cues is captured by an inter-cue correlation matrix. Here, the extremity of cues can only be learnt through repeated experience with the cue values, and the "extreme cues" may not appear to be outliers by the standards in Harries et al. (2004). Future studies should attempt to narrow the gap between research on multi-cue judgments and judge-advisor systems, and to use insights from multi-cue judgment to understand the cognitive and statistical underpinnings of judges' advice aggregation processes.

Despite our finding that many participants used the structure of cue correlations to determine cue weights, the equal weights rule was the best model for a majority of participants in all our studies. Some of these participants might have failed to learn the cue structure from the Tasks 1 and 2 and believed that the cues were equally correlated to each other. For these participants the $\boldsymbol{\beta_{EV1}}$ model that minimizes the maximum risk could not be distinguished from the $\boldsymbol{\beta_{EW}}$ model. In a sense, our current analysis could be considered as a conservative evaluation of the use of the $\boldsymbol{\beta_{EV1}}$ model.

Note that a post hoc model recovery study suggested that there may have been some limitations to our ability to adequately

discriminate the β_{EV1} and β_{EW} models in Study 1 and the treatment condition of Study 2. Particularly, in these studies, the models make somewhat similar predictions in Task 3, implying that some β_{EV1} decision makers could have been incorrectly determined to be β_{EW} decision makers and that some β_{EW} decision makers could have been incorrectly determined to be β_{EV1} decision makers. Note that this issue is primarily the product of the statistical structure of the cue space: Whenever one cue is more correlated with remaining cues than the remaining cues are with each other, a decision rule that places a high weight on the special cue will yield predictions that are close to the overall average of all the cues. As our stimuli in Task 3 were representative of the cue distribution in consideration, this issue was unavoidable. Thus, in future work, it is necessary to design experimental stimuli that are better suited for disentangling the different candidate judgment models. This involves using specialized cue distributions in the criterion judgment task (cue distributions that do not appropriately mimic the underlying cue-correlation structure). Such stimuli will also enable researchers to study other intriguing questions arising in our experimental setting. For example, how do cue integration strategies change over time when decision makers know inter-cue correlations but not cue-criterion relationships? In our current study, we assume that individuals use a common strategy in all trials, as the current stimuli design is not sensitive enough for reliably detecting changes in weighting schemes.

Relatedly, although our three studies show that the β_{EV1} model describes the behavior of some participants, we cannot make the causal claim that these participants are behaving in the manner predicted by β_{EV1} because they optimally decompose and utilize the cue correlation matrix. In other words, our model is purely descriptive. Future work is needed to better understand the causes behind the observed behavioral patterns. As a starting point, the judge-advisor systems literature provides a social explanation: Decision makers may treat advisors underweighted by the β_{EV1} model as more extreme and thus rely less on their recommendations (Harries et al., 2004; Yaniv, 1997). Relatedly, participants might assume that advisors that are highly correlated with others are experts and thus trust their advice more (e.g., Birnbaum, Wong, & Wong, 1976; Sniezek & Van Swol, 2001). To further test this explanation, researchers can manipulate prior experiences with the advisors and examine how this alters the influence of the cue structure. The judge-advisor systems literature may also shed light on domain differences in cue integration. In this paper, our tasks involve predicting stock price (Studies 1 and 2) or restaurant quality (Study 3). We implicitly assume that advice is integrated in a similar manner in both domains; however, prior work has found that people use social cues differently in different settings (Dalal & Bonaccio, 2010; Van Swol, 2011). Future studies should systematically examine how cue structures influence advice integration in different domains, and whether cue structures learnt in one domain generalize to other domains (i.e., interdomain extrapolation).

On the cognitive side, participants may be using other heuristics, not considered in the current paper, that in certain cases mimic the behavior of the β_{EV1} model. For example, as suggested by one of our reviewers, the mean correlation with other cues is higher for Cue 1, relative to Cues 2–4 in the inter-cue correlation structure presented in Figure 1a. It is thus possible that participants might have integrated cues according to their mean correlations with other cues, instead of following the β_{EV1} model. More generally, the behaviors observed in this study could emerge due to a variety of cognitive mechanisms, which are not always optimal or explicitly rational. Better understanding these mechanisms is a useful topic for future work. Future work should also attempt to integrate the insights of cognitive models of multi-cue judgment, such as those relying on neural network representations (Glöckner, Hilbig, & Jekel, 2014) or exemplar memory-based (Juslin et al., 2008). Such models have not been applied to settings in which cue-criterion relationships are unknown. However, they nonetheless provide formal predictions regarding the learning and representation of cue knowledge and its relationship with the statistical structure of the judgment environment. For this reason, they may provide a more adequate framework for understanding the cognitive underpinnings of the optimal improper linear model in the absence of cue-criterion knowledge.

In conclusion, our studies provided empirical evidence regarding the improper linear models decision makers use when the cue structure is known but the cue-validities are unknown. Our work shows that some decision makers overweigh a subset of cues compared with the remaining cues, even though they have no information regarding these cues' validities. One important implication of this finding is that not only can the use of improper linear models be a result of overweighting cues with higher validities (as found in studies with known cue-criterion relationships), but it may also arise naturally from the cue structure learnt by the decision maker (as in our study, with unknown cue-criterion relationships). Moreover, these naturally arising improper linear models have compelling optimality properties.

ORCID

Wenjia Joyce Zhao https://orcid.org/0000-0003-1771-6462

REFERENCES

Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, *21*(1), 243–247. https://doi.org/10.1007/BF02532251

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum.

Armelius, K., & Armelius, B.-Å. (1976). The effect of cue-criterion correlations, cue intercorrelations and the sign of the cue intercorrelation on performance in suppressor variable tasks. *Organizational Behavior and Human Performance*, *17*(2), 241–250. https://doi.org/10.1016/0030-5073(76)90065-9

Armstrong, J. S. (2001). *Principles of forecasting: A handbook for researchers and practitioners*. Dordrecht, Netherlands: Kluwer.

Azen, R., & Budescu, D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods*, *8*(2), 129–148. https://doi.org/10.1037/1082-989X.8.2.129

Bhatia, S. (2018). Decision making in environments with non-independent dimensions. *Journal of Behavioral Decision Making*, *31*(2), 294–308. https://doi.org/10.1002/bdm.1964

Birnbaum, M. H., Wong, R., & Wong, L. K. (1976). Combining information from sources that vary in credibility. *Memory and Cognition*, *4*, 330–336. https://doi.org/10.3758/BF03213185

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. https://doi.org/10.1016/j.obhdp.2006.07.001

Broomell, S. B., & Budescu, D. V. (2009). Why are experts correlated? Decomposing correlations between judges. *Psychometrika*, *74*(3), 531–553. https://doi.org/10.1007/s11336-009-9118-z

Brunswik, E. (1952). *The conceptual framework of psychology*. *International Encyclopedia of Unified Science*, *Vol. 1*, *No. 1*. Chicago: The University of Chicago Press.

Budescu, D. V., & Rantilla, A. K. (2000). Confidence in aggregation of expert opinions. *Acta Psychologica*, *104*(3), 371–398. https://doi.org/10.1016/S0001-6918(00)00037-8

Budescu, D. V., Rantilla, A. K., Yu, H.-T., & Karelitz, T. M. (2003). The effects of asymmetry among advisors on the aggregation of their opinions. *Organizational Behavior and Human Decision Processes*, *90*(1), 178–194. https://doi.org/10.1016/S0749-5978(02)00516-2

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583. https://doi.org/10.1016/0169-2070(89)90012-5

Czerlinski, J., Gigerenzer, G., & Goldstein, D. G. (1999). How good are simple heuristics? In G. Gigerenzer, P. M. Todd, & the ABC Research Group (Eds.), *Simple heuristics that make use smart* (pp. 97–118). New York, NY: Oxford University Press.

Dalal, R. S., & Bonaccio, S. (2010). What types of advice do decision-makers prefer? *Organizational Behavior and Human Decision Processes*, *112*(1), 11–23. https://doi.org/10.1016/j.obhdp.2009.11.007

Dana, J., & Davis-Stober, C. P. (2016). Rational foundations for fast and frugal heuristics: An improper linear models approach. *Minds and Machines*, *26*, 61–86. https://doi.org/10.1007/s11023-015-9372-z

Davis-Stober, C. P. (2011). A geometric analysis of when fixed weighting schemes will outperform ordinary least squares. *Psychometrika*, *76*, 650–669. https://doi.org/10.1007/s11336-011-9229-1

Davis-Stober, C. P., Dana, J., & Budescu, D. V. (2010a). A constrained linear estimator for multiple regression. *Psychometrika*, *75*(3), 521–541. https://doi.org/10.1007/s11336-010-9162-8

Davis-Stober, C. P., Dana, J., & Budescu, D. V. (2010b). Why recognition is rational: Optimality results on single-variable decision rules. *Judgment and Decision making*, *5*(4), 216.

Dawes, R., Faust, D., & Meehl, P. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668–1674. https://doi.org/10.1126/science.2648573

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*(7), 571–582. https://doi.org/10.1037/0003-066X.34.7.571

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, *53*(2), 134–140. https://doi.org/10.1037/h0045156

Fischer, I., & Harvey, N. (1999). Combining forecasts: What information do judges need to outperform the simple average? *International Journal of Forecasting*, *15*(3), 227–246. https://doi.org/10.1016/S0169-2070(98)00073-9

Fishburn, P. C. (1974). Exceptional paper—Lexicographic orders, utilities and decision rules: A survey. *Management Science*, *20*(11), 1442–1471. https://doi.org/10.1287/mnsc.20.11.1442

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology*, *62*(1), 451–482. https://doi.org/10.1146/annurev-psych-120709-145346

Gigerenzer, G., Todd, P. M., & ABC Research Group (1999). *Simple heuristics that make us smart*. Oxford, England: Oxford University Press.

Glöckner, A., Hilbig, B. E., & Jekel, M. (2014). What is adaptive about adaptive decision making? A parallel constraint satisfaction account. *Cognition*, *133*(3), 641–666. https://doi.org/10.1016/j.cognition.2014.08.017

Griffiths, T. L., & Tenenbaum, J. B. (2016). Optimal predictions in everyday cognition. *Psychological Science*, *17*(9), 767–773.

Harries, C., Yaniv, I., & Harvey, N. (2004). Combining advice: The weight of a dissenting opinion in the consensus. *Journal of Behavioral Decision Making*, *17*(5), 333–348. https://doi.org/10.1002/bdm.474

Juslin, P., Karlsson, L., & Olsson, H. (2008). Information integration in multiple cue judgment: A division of labor hypothesis. *Cognition*, *106*(1), 259–298. https://doi.org/10.1016/j.cognition.2007.02.003

Keeney, R. L., & Raiffa, H. (1993). *Decisions with Multiple Objectives*. Cambridge, England: Cambridge University Press.

Klayman, J. (1988). Cue discovery in probabilistic environments: Uncertainty and experimentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(2), 317–330.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling (quantitative applications in the social sciences)*. Beverly Hills, CA: Sage Publications.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores.

Maines, L. A. (1996). An experimental examination of subjective forecast combination. *International Journal of Forecasting*, *12*(2), 223–233. https://doi.org/10.1016/0169-2070(95)00623-0

Marewski, J. N., & Mehlhorn, K. (2011). Using the ACT-R architecture to specify 39 quantitative process models of decision making. *Judgment and Decision Making*.

Martignon, L., & Hoffrage, U. (2002). Fast, frugal, and fit: Simple heuristics for paired comparison. *Theory and Decision*, *52*, 29–71. https://doi.org/10.1023/A:1015516217425

Newell, B., & Bröder, A. (2008). Cognitive processes, models and metaphors in decision research. *Judgment and Decision making*, *3*(3), 195.

Oaksford, M., & Chater, N. (2007). *Bayesian rationality*Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198524496.001.0001

Oppenheimer, D. M., & Kelso, E. (2015). Information processing as a paradigm for decision making. *Annual Review of Psychology*, *66*, 277–294. https://doi.org/10.1146/annurev-psych-010814-015148

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*Cambridge University Press. https://doi.org/10.1017/CBO9781139173933

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111. https://doi.org/10.2307/271063

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2013). Learning hierarchical category structure in deep neural networks. I*n Proceedings of the 35th annual meeting of the Cognitive Science Society* (pp. 1271–1276).

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Sniezek, J. A., & Buckley, T. (1995). Cueing and cognitive conflict in judge-advisor decision making. *Organizational Behavior and Human Decision Processes*, *62*(2), 159–174. https://doi.org/10.1006/obhd.1995.1040

Sniezek, J. A., & Van Swol, L. M. (2001). Trust, confidence, and expertise in a judge-advisor system. *Organizational Behavior and Human Decision Processes*, *84*(2), 288–307. https://doi.org/10.1006/obhd.2000.2926

Soll, J. B. (1999). Intuitive theories of information: Beliefs about the value of redundancy. *Cognitive Psychology*, *38*(2), 317–346. https://doi.org/10.1006/cogp.1998.0699

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318. https://doi.org/10.1016/j.tics.2006.05.009

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Van Swol, L. M. (2011). Forecasting another's enjoyment versus giving the right answer: Trust, shared values, task effects, and confidence in improving the acceptance of advice. *International Journal of Forecasting*, *27*(1), 103–120. https://doi.org/10.1016/j.ijforecast.2010.03.002

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196. https://doi.org/10.3758/BF03206482

Weber, E. U., & Johnson, E. J. (2009). Mindful judgment and decision making. *Annual Review of Psychology*, *60*, 53–85. https://doi.org/10.1146/annurev.psych.60.110707.163633

Yaniv, I. (1997). Weighting and trimming: Heuristics for aggregating judgments under uncertainty. *Organizational Behavior and Human Decision Processes*, *69*, 237–249. https://doi.org/10.1006/obhd.1997.2685

Yaniv, I. (2004). The benefit of additional opinions. *Current Directions in Psychological Science*, *13*(2), 75–78. https://doi.org/10.1111/j.0963-7214.2004.00278.x

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Zhao WJ, Davis-Stober CP, Bhatia S. Optimal cue aggregation in the absence of criterion knowledge. *J Behav Dec Making*. 2019;1–16. https://doi.org/10.1002/bdm.2123