Does the aggregation of multiple forecasts (made at different time points) improve the prediction accuracy compared to the use of the most recent single forecast?

Martino Avalle

s5164257

Department of Psychology, University of Groningen

PSB3E-BT15: Bachelor Thesis

Group 26

Supervisor: Fridtjof Petersen

Second evaluator: Dr. Oliver Weigelt

In collaboration with: Fabienne Liepelt, Marieke Boerendonk, Martino Avalle, Miora Haslacher.

July 2nd, 2025

Abstract

Past research has explored the concept of aggregation in the context of human forecasting. Specifically, within-person aggregation has been found to be advantageous in cases where only singular sources are available. This process works according to the idea that averaging forecasts from the same person at different times can reduce random error. In the current paper, we try to translate this idea to the field of affective forecasting. We hypothesized that aggregating the two previous predictions for any time point would offer a predictive advantage compared to just using the single most recent forecast. Our participants repeatedly rated their current affect and predicted the emotional levels they would feel in 3 and 6 hours for 5 times a day for 14 days (n=30). These predictions were given for four emotions along the dimensions of positive and negative affect. Absolute errors were calculated, and our results showed no significant difference between using aggregation and using the single most recent forecast. The results were similar and non-significant for both positive and negative affect. Our findings are, thus, consistent with the idea that the single most recent forecast is generally more accurate than the previous ones, and contrast, within the boundaries of our experiment, with the idea that aggregation could be useful to diminish predictive error. We suggest that this could be because affective forecasting is mainly driven by systematic biases rather than random noise, which aggregation does not cancel.

Keywords: aggregation, affective forecasting, time, wisdom of the inner crowd, affect, accuracy, prediction error

Introduction

Whether you are deciding whether to buy an expensive snack or end a relationship, considering how decisions will affect your feelings is a recurrent experience in our daily lives. As defined by Wilson and Gilbert (2003), "predicting one's emotional reactions to future events, including the intensity, duration, and valence (positive or negative) of those emotions" is referred to as affective forecasting. This is seen as a universal human experience fundamental for emotional management and planning (Gilbert & Wilson, 2007). Predicting our mood is something that we do every day, even though we are not always consciously aware of it (Wilson & Gilbert, 2003).

Research in the field consistently demonstrates the limited accuracy of affective forecasting (Wilson & Gilbert, 2003). Specifically, individuals often misjudge the intensity and duration of their future emotions, leading to systematic errors in their predictions (Dunn et al., 2007). These inaccuracies usually stem from cognitive biases, memory distortions, and the tendency to underestimate the psychological mechanisms that regulate emotions (Wilson et al., 2001). For example, *focalism*, the tendency to overemphasize the focal event and neglect other influences on future emotions (Wilson et al., 2000), is one of the main causes of errors explored until now. Another one is *immune neglect*, which refers to the underestimation of psychological coping mechanisms that mitigate negative emotional reactions (Gilbert et al., 1998). In addition to these biases that are experienced by everyone, individual differences, such as personality traits and emotional intelligence, influence forecasting ability, with some people more prone to misjudgment than others (Hoerger et al., 2012).

This general inaccuracy in foreseeing future well-being can have significant negative implications across various domains. Research in consumer behavior found that biased predictions and the biased recall typical of affective forecasting can have negative consequences on consumers' spending behaviors. In their study, Pollai et al. (2009) performed a two-wave longitudinal survey on 86 consumers in the UK and Austria who had just bought a pair of shoes. They assessed participants' current consumption-related positive emotions and predictions regarding their feelings in the next weeks. Two or four weeks (based on the condition they were in) after the initial survey, participants were then asked to report their current emotions about the purchase and recall their original prediction. Results showed that participants overestimated the decrease in consumption-related positive emotions over the interval between the two surveys, as well as exaggerated this decrease in their recollections. These results are relevant because a perceived large decrease in positive emotions might induce people to invest in a new product and thus spend more money in an attempt to compensate for the decreased emotions (Pollai et al., 2009), leading to unnecessary expenses.

Another compelling example of the implications of affective forecasting inaccuracy is the medical field, where these errors can often lead to unfortunate consequences. In their study, Halpern and Arnold (2008) analysed three cases to illustrate that even competent and educated individuals can end up making health decisions seen as tragic by their clinicians, when driven by their inaccurate beliefs. Psychological research consistently demonstrates that people poorly predict their future ability to adapt to adversity, often basing these predictions on inaccurate and non-realistic beliefs(Gilbert et al., 1998; Halpern & Arnold, 2008; Ubel et al., 2005). This inability to correctly envision future emotional states and adjust to future scenarios can prevent individuals from making important health decisions that could improve their health condition or emotional well-being (Halpern & Arnold, 2008)

These are only some examples of the implications of these errors on people's lives, yet they highlight the negative impact that forecasting inaccuracy can have on people's lives. Such concerns strongly point to a need to improve our understanding of why people's predictions can be so inaccurate, and to use this understanding to improve emotional forecasts. One particularly important contextual factor that can be investigated to understand people's poor predictions is time.

Past studies have highlighted the importance of the concept of time in the context of human and affective forecasting (Buehler & McFarland, 2001). Specifically, both the temporal course and the distance between prediction and event have been investigated as influences on forecasting accuracy (Finkenauer et al., 2007). Past research is consistent with the idea that predictions for events closer in time are more accurate than ones for events that are more distant(Liberman et al., 2002). The main reason for this would be that closer futures are constructed in more concrete terms than how the further ones are, although this was studied for events that largely differ in time (1 day vs 1 year)(Liberman et al., 2002). The study by Finkenauer et al. (2007), who specifically investigated the role of time in affective forecasting, found a nuanced temporal effect partially consistent with this idea. They analyzed participants' predictions regarding their positive and negative affect after a driving license exam. What they found is that participants were more accurate at predicting positive affect when the focal event was more distant, and more accurate at predicting negative affect when it was closer. Supposedly, this would be because when the event is distant people down-regulate positive affect and up-regulate negative affect in order to maintain motivation, while when the event gets closer people tend to do the opposite in order to mitigate the negative impact of the event in case of a negative outcome (Finkenauer et al., 2007)

These findings suggest that time would be a determinant of systematic biases in affective forecasting. These time-related errors underscore a fundamental challenge in understanding and improving affective forecasting. We propose that forecast aggregation could be one way to account for and potentially diminish these kinds of errors. In the context of human forecasting, forecast aggregation leverages the "wisdom of the crowd" by combining multiple individual predictions into a single estimate, an approach that often surpasses the accuracy of most standalone judgments. Several studies, including Mellers et al. (2023), demonstrate that aggregated forecasts outperform most individual judgments. This process typically involves averaging predictions from a selected group or the entire set of forecasters. A good example of this principle is presented by Codi et al. (2022). In their study, they analyzed predictions from human judgment surveys regarding COVID-19's weekly incident cases, hospitalizations, deaths, vaccinations, cumulative first doses of vaccines, and prevalence of immunity-evading variants. A remarkable finding is that consensus predictions for weekly incident deaths were closer to the truth than 75% of individual predictions, while in one of their surveys, the aggregated prediction was closer to the truth than any single individual one.

However, this approach remains untested in affective forecasting, where predictions involve highly personal, subjective emotional states. The combination of forecasts from different sources, in this domain, is not feasible due to the inherent individual-specificity of affective forecasts (e.g., predicting one's own affective reaction to a future relationship break-up). Thus, standard crowd aggregation becomes largely impractical. A different approach is needed to translate the benefits of aggregation to this domain. Building on the evidence provided by

Finkenauer et al. (2007) on the accuracy shifts related to temporal distance for negative and positive affect, we propose that an aggregation method based on combining forecasts made by single individuals at different time points might be more suitable for this domain.

Although never tested in the realm of affective forecasting, the “wisdom of the inner crowd” is not a new concept. Remarkably, this has been observed across various contexts, such as percentage estimation, general knowledge questions, date estimation, and quantity estimation (Vul & Pashler, 2008). For example, van Dolder and van den Assem (2017) analyzed the results from three promotional events over three years, organized by the Dutch state-owned Holland Casino. People were asked to estimate the exact number of objects in a transparent container, which had the same shape and contained the same number of objects in each location. By predicting the believed right number of objects in the container, they had the possibility to win 100.000 euros. This estimation could be repeated every time they visit one of the casinos. Their results showed that within-person aggregation offered an advantage in prediction accuracy compared to individual predictions, although this was not comparable to the advantage offered by between-person aggregation. In their conclusion, they suggested that this kind of aggregation could potentially be useful in situations where only one individual can make sufficiently informed estimates (van Dolder & van den Assem, 2017). Overall, it is difficult to say whether this effect will translate to the domain of affective forecasting. In case a translation to this domain was to be found, it would open new avenues for research into improving the accuracy of emotional predictions.

The reasoning behind the mechanism of aggregation stems from the idea that when people are asked to make a prediction, their brain holds a range of diverse, sometimes conflicting information regarding the same object or envisioned time point. When the prediction is performed, they take a subsample from this set of information, which can introduce a bit of random error (Litvinova et al., 2019). When asked to make a second prediction, assuming it is not biased by the first one, they will produce a slightly or sometimes drastically different prediction, with a different random noise (Litvinova et al., 2019). Accordingly, aggregating will average out this random error, refining the final estimate (Vul & Pashler, 2008). The benefits of aggregation are maximized when a significant temporal interval is introduced between predictions, increasing their independence and thus contributing to a greater reduction in error (Fiechter & Kornell, 2021; Vul & Pashler, 2008).

In sum, the introduction of aggregation might be a useful addition to reduce the negative impact of momentary biases and situational influences, thereby decreasing the number of sources of errors present in the equation. This notion is supported by Takano and Ehring's (2024) finding that current affective states have a particularly heavy impact on people's daily predictions in affective forecasting. By combining forecasts from different time points, these momentary fluctuations may cancel each other out, leading to more balanced and accurate predictions. This parallels findings from the broader forecasting literature, where aggregation tends to outperform individual estimates by minimizing the impact of extreme or erroneous judgments (Armstrong, 2001). However, affective forecasting is mainly driven by unidirectional biases rather than random noise. The main reasons for this difference are the impact bias (the overestimation of how strongly and long we will feel a certain way) and immune neglect(the failure to anticipate coping mechanisms). Since these mechanisms all pull in the same direction, causing people to overestimate their future emotions, it is possible that the "cancellation" mechanism leveraged by the "wisdom of the inner crowd" effect might not effectively operate in this domain.

Accordingly, aggregating two overly positive estimates might reinforce the bias rather than average it out.

As of the current moment the literature on affective forecasting has investigated different themes, ranging from gaining an understanding of the different biases underlying forecasting error (Wilson & Gilbert, 2003), to researching on ideating and testing of novel interventions to diminish inaccuracy and prevent its negative consequences (Ellis et al., 2018). Specifically, in relation to the context of our exploration, the study by Finkenauer et al. (2007) is the only concrete effort we could find studying the relationship between time and forecasting accuracy, although their design emphasized a specific focal event rather than everyday mood forecasts. Various papers have also succeeded in creating useful interventions to improve forecasting accuracy (Hoerger et al., 2010; Walsh & Ayton, 2009); however, we found no study using time-related variables to diminish prediction error. Accordingly, in line with the findings on the "wisdom of the inner crowd" in the field of human forecasting (Fiechter & Kornell, 2021; van Dolder & van den Assem, 2017; Vul & Pashler, 2008), the current study tries to transfer forecasts aggregation to the prediction of emotional states in the close future, suggesting that averaging different predictions might have a positive impact on the accuracy of the final prediction.

The main hypothesis of this study is that aggregating the two predictions previous to a single time point will yield more accurate results than the single most recent prediction. We also acknowledge that a nuance for the valence of the predicted emotion could be present. In fact, building on evidence from Finkenauer et al. (2007), we suggest that aggregation could yield more accurate results against the most recent single prediction, especially for positive rather than negative affect. The present study is part of a bigger study aimed at conceptually replicating

Takano and Ehring's study, where the effect of accounting for uncertainty on improving the accuracy of affective forecasting is investigated.

To test our hypothesis, we collected data using the Experience Sampling Method (ESM) (Csikszentmihalyi et al., 1977), where participants are prompted multiple times a day to fill out the same questionnaire repeatedly. Specifically, participants will be prompted about how they feel at the current moment, as well as how they will feel at the next two time points, which are 3 and 6 hours in the future. The predictions made for the different time intervals will then be aggregated, leveraging the "wisdom of the inner crowd" effect. We decided to adopt the ESM to collect real-time data across multiple time points, enabling us to capture the dynamic nature of emotional states and predictions (Csikszentmihalyi & Larson, 1987). Additionally, the use of this method serves to largely reduce recall bias, a limitation commonly found in self-report studies (Fritz et al, 2024).

We aim with this study to offer a new perspective on affective forecasting and highlight the potential of novel methodological approaches in understanding and improving forecasting accuracy.

Method

Participants

This research project is a quantitative observational study about affective forecasting. Data was obtained via ESM, through which participants can predict their future emotions and report their real-time emotional experiences on their mobile device. By comparing these predictions to the emotions they later reported experiencing, we measured prediction accuracy or prediction error (i.e, how well their predicted emotions matched their actual emotions). Additionally, point and interval prediction were used to assess the impact of uncertainty on self-ratings and predicted emotions (Takano & Ehring, 2023), but this was not relevant for the current paper. Ethical approval was obtained through the FMG research lab (FMG-12534_2025). As we aimed to conceptually replicate and extend the study of Takano and Ehring (2023), we aimed for a sample size of 68. The final sample consisted of 30 first-year psychology students from the University of Amsterdam (20 women, 10 men) between X and Y years old (*Mean* age=19.97; *SD* age=1.83). Individuals provided their informed consent before participating further.

Procedure

Data collection methods

To be eligible for this research project, students had to own a smartphone, understand the English language, and not have been diagnosed with depression or anxiety. The study was advertised through flyers on campus, social media, and via the student participation tool. Therefore, we used a convenience sample for our study. Students receive course credit for participating.

Consent was received through a SONA Qualtrics questionnaire, which also included instructions, and upon enrollment, participants received a link to the m-Path questionnaire via email.

Data were collected using ESM via the m-Path app (Mestdagh et al., 2023), which enabled participants to complete short questionnaires on their phones while engaging in their daily activities. The ESM questionnaire focused on four core emotions: happiness, relaxation, sadness, and anxiety. For each emotion, participants responded to a set of items that assessed their current emotional state and were asked to predict how they will feel in the coming hours. Each of these sets, consisting of three items, asked the participants to rate the intensity of emotion on a continuous scale. Additionally, participants defined a confidence range, indicating the minimum and maximum intensity values in which they were 95% certain their emotions fell.

Participants received five prompts per day (9 am, 12 pm, 3 pm, 6 pm, 9 pm) over a 14-day period to complete the ESM questionnaire. The questionnaire takes around 5-10 minutes to complete. Upon receiving a notification, participants had a 30-minute time window to respond before the prompt expired, with a reminder sent after 15 minutes. This time limit was put in place to ensure that the time between questionnaires remains great enough to make affective forecasts for the next beep. The order of the items for each emotion was the same for every participant, but the order of the emotions was randomised for every time point for each participant.

Materials

The self-report instruments consisted of prompts assessing participants' current levels of four emotional states: happiness, relaxation, anxiety, and sadness. The choice of the emotions to test for in our experiment came as a direct consequence of the decision to replicate the study conducted by Takano and Ehring (2024), where sadness and anxiety were mapped onto the dimension of negative affect, and happiness and relaxation were mapped onto that of positive affect. Additionally, these emotions easily fit into the generally recognized dimensions of valence (direction of emotion) and arousal (level of activation)(Mathieu & Gosling, 2012; Russell, 2003).

At each assessment point, participants were asked to provide two types of forecasts for each emotional state: a point prediction (a single value on a 1-100 scale) and an interval prediction (a range defined by a lower and upper bound on the 1-100 scale). Both were measured using a Visual Analogue Scale (VAS) where the starting point (1) was anchored as “not at all” and the end point (100) as “extremely”. These predictions were asked for three distinct time points: the current moment ("Please rate how much you experience [emotion] at the moment"), 3 hours into the future, and 6 hours into the future ("Please estimate how much you will experience [emotion] at [time of next beep]."). For the interval predictions, participants provided two credible bounds on the same scale ("I am 95% sure that my [emotion] level is at least [at most]..."; “I am 95% sure that my [emotion] level will be at least [at most] at [time of next beep].").

Data Preprocessing

Firstly, data preprocessing was done in R Statistical Software (v4.3.3; R Core Team, 2023) and excluded participants with less than 30% compliance (i.e., 21 beeps) from the dataset to ensure that our data is in line with the data from Takano and Ehring (2024). The average completion rate of the questionnaires in our dataset is 70% (*SD* = 19%), with an average compliance range of 29% to 94%. Moreover, 29 of the total number of participants (30) filled out 21 or more questionnaires. It is important to note that participants could only fill out the questionnaire once and thus were not able to adjust their answers once they had submitted their response.

As the next step in preparing the data, we computed aggregated predictions to investigate the main hypothesis explored by this paper. For each participant, these were calculated by averaging the predictions produced 3 and 6 hours ahead of each time point. This approach was taken to compare the aggregated predictions with the most recent forecast for each assessed time point. This process was conducted for both positive (PA) and negative affect (NA). Subsequently, we calculated absolute errors by subtracting either the most recent forecast or the computed aggregated forecast scores from the actual reported scores for each time point, obtaining the absolute difference for both affective dimensions.

Data analysis

Data analysis was conducted using R Statistical Software (v4.3.3; R Core Team, 2023) and JASP 0.18.3 (JASP Team, 2024).

To examine differences in absolute error across the prediction methods, a linear mixed model(Bates et al., 2015; Brown, 2021; Kuznetsova et al., 2017) was employed using R (v4.3.3; R Core Team, 2023). These types of models analyze observations that are grouped, enabling researchers to account for population-level effects and subject-specific variability. The dependent variable was absolute prediction error. The independent variable, representing whether the prediction was obtained by using the most recent forecast or by aggregating the forecasts obtained 3 and 6 hours ahead of each time point, was included as the only fixed factor. Finally, Participant ID was included as a random factor to account for the interdependence of observations from the same individuals, with random intercepts to allow each participant to retain their own baseline level of prediction error.

Results

Descriptive statistics regarding the error for the most recent vs aggregated prediction are presented in Table 1 for positive affect and Table 2 for negative affect. For positive affect, the most recent forecasts show a slightly higher mean error (M=11.667, SD=10.521) than the aggregated ones (M=11.222, SD=10.088). A very similar pattern was found for negative affect, where the most recent forecasts exhibited a slightly higher mean error(M=11.234, SD=10.716) than the aggregated counterpart(M=10.889, SD=10.527).

Table 1

Descriptive statistics for the absolute prediction error for positive affect.

	error_PA
Mean	11.444
Std. Deviation	10.306
Minimum	0.000
Maximum	72.000

Table 2

Descriptive statistics for the absolute prediction error for negative affect.

	error_NA
Mean	11.062
Std. Deviation	10.621
Minimum	0.000
Maximum	81.500

A significant positive correlation was found between the errors for positive and negative affect. Its size was moderate (r=0.350, p<.001) (Table 3), meaning that when the error for positive affect increased, it increased for negative affect as well.

Table 3

Correlation table for the relationship between the error for positive and negative affect.

Variable		error_NA	error_PA
1. error_NA	*Pearson's r*	—	
	p-value	—	
2. error_PA	*Pearson's r*	0.350	—
	p-value	< .001	—

A linear mixed model(LMM) was employed to examine the effect of prediction type (aggregated vs. most recent) on affective forecasting accuracy. To account for the nested structure of the data (multiple predictions per participant), a random intercept for participant was included in the model. Two separate analyses were conducted for positive and negative affect, respectively. We conducted a posterior predictive check and tested homoscedasticity, normality of residuals, and of random effects for both positive and negative affect. For both affective dimensions, the posterior predictive check, homoscedasticity, and normality of residuals were violated, while the normality of random effects was not, as shown in Figures 1 and 2 (Appendix B). Due to violations of these assumptions for the LMM, generalized linear mixed models (GLMM) (Ng & Cribbie, 2016) were subsequently performed, for which all assumptions were met as shown in Figure 3 and Figure 4 (*Appendix B*). Most importantly, the results obtained from both the LMM and GLMM analyses were consistent, yielding substantively similar conclusions.

Despite the slight difference suggested by the descriptive statistics, the model fitted for positive affect indicated no significant main effect of the prediction type on the dependent variable absolute prediction error ($F(1,1854.87)=0.541$, $p=0.462$)(Table 4). The estimated marginal mean for the aggregated predictions was 10.727 ($SE=0.669$, 95% CI [9.416,12.038]), while the one for the most recent predictions was 11.072 ($SE=0.669$, 95% CI [9.760,12.383]). These means are consistent with the non-significant main effect, as their confidence intervals largely overlap.

Table 4

ANOVA summary LMM positive affect

Effect	***df***	***F***	***p***
type	1, 1852	0.913	0.339

Note. Model terms tested with Satterthwaite testMethod.

Note. The following variable is used as a random effects grouping factor: 'ParticipantID'.

Note. Type III Sum of Squares

Table 5

Model fit LMM positive affect

Deviance (REML)	*log Lik.*	*df*	*AIC*	*BIC*
14082.483	-7041.241	4	14090.483	14112.643

Deviance (REML)	*log Lik.*	*df*	*AIC*	*BIC*

Note. The model was fitted using restricted maximum likelihood. Please note that models with different fixed effects cannot be compared when REML is used. To use ML, switch 'Test method' to 'Likelihood ratio tests'.

Table 6

Fixed Effects Estimates LMM positive affect

Term	*Estimate*	*SE*	*df*	*t*	*p*
Intercept	11.601	0.474	25	24.478	< .001
type (1)	-0.223	0.233	1852	-0.956	0.339

Note. The intercept corresponds to the (unweighted) grand mean; for each factor with k levels, k - 1 parameters are estimated with sum contrast coding. Consequently, the estimates cannot be directly mapped to factor levels. Use estimated marginal means for obtaining estimates for each factor level/design cell or their differences.

Similarly, a different linear mixed-effects model was conducted to test the hypothesis for the variable negative affect. This model showed, once again, no significant main effect (F(1,1852.84)=0.913, p=0.339)(Table 7). The estimated intercept(most recent forecast), was 11.601 (SE=0.474), which was significantly different from zero (t=24.478, p<.001)(Table 9). The estimated marginal mean for the aggregated predictions was 11.378 (SE=0.528, 95% CI

[10.343,12.413]), and for the most recent predictions was 11.823 (*SE*=0.528, 95% CI [10.788,12.858]).

Table 7

ANOVA summary LMM negative affect

Effect	*df*	*F*	*p*
type	1, 1854	*0*.541	0.462

Note. Model terms tested with Satterthwaite testMethod.

Note. The following variable is used as a random effects grouping factor: 'ParticipantID'.

Note. Type III Sum of Squares

Table 8

Model fit LMM negative affect

Fit statistics

Deviance (REML)	*log Lik.*	*df*	*AIC*	*BIC*
14116.907	-7058.453	4	14124.907	14147.067

Note. The model was fitted using restricted maximum likelihood. Please note that models with different fixed effects cannot be compared when REML is used. To use ML, switch 'Test method' to 'Likelihood ratio tests'.

Table 9

Fixed Effects Estimates LMM negative affect

Term	*Estimate*	*SE*	*df*	*t*	*p*
Intercept	10.899	0.627	29.716	17.392	< .001
type (1)	-0.172	0.234	1854.872	-0.736	0.462

Note. The intercept corresponds to the (unweighted) grand mean; for each factor with k levels, k - 1 parameters are estimated with sum contrast coding. Consequently, the estimates cannot be directly mapped to factor levels. Use estimated marginal means for obtaining estimates for each factor level/design cell or their differences.

An exploratory analysis was performed to compare the accuracy of two-step-ahead predictions (oldest) with that of one-step-ahead predictions (most recent). The absolute error was calculated using the same procedure as previously employed for the main analysis. The comparison was then conducted through an LMM, which revealed a significant difference between the two estimates for both positive (F(1,99.59)=6.646, p=0.011) and negative affect (F(1,197.42)=4.616, p=0.033). In particular, for both dimensions, higher accuracy scores for the most recent predictions were found.

Discussion

Research from other fields suggests that aggregating different forecasts from various sources may be a mechanism to improve forecasting accuracy, thereby reducing the discrepancy between predictions and real-life outcomes. On top of this, research on human forecasting has previously investigated the so-called “wisdom of the inner crowd”, or within-person aggregation. This phenomenon suggests that increases in accuracy can be obtained by averaging predictions produced by a single individual at different time points. The idea for this paper started with the suggestion that this concept might be translated to affective forecasting.

Interpretation of results

The primary aim of our study was to investigate whether aggregating predictions would lead to greater accuracy than relying solely on the most recent prediction within the domain of affective forecasting.

Contrary to our hypotheses and the general trend observed in research on within-person aggregation outside of the affective forecasting domain (Fiechter & Kornell, 2021; van Dolder & van den Assem, 2017; Vul & Pashler, 2008), our analysis revealed no statistically significant advantage for aggregating predictions over the most recent prediction. This indicates that, within the boundaries of the current study, the predictive utility of aggregating past emotional predictions did not significantly outperform a simpler "single most recent prediction" approach.

One potential reason for the current reported absence of a predictive advantage in using aggregation could be related to the nature of the data being aggregated. Accordingly, unlike the objective quantities considered in human judgment (e.g., prices, demand, temperature), affective states are subjective and context-dependent. In this regard, a prediction might quickly lose relevance following a change in internal states or the unfolding of a new event. This suggests that the predictive value of older forecasts could fluctuate over time and often diminish quickly, making the use of aggregation less effective than using the closest prediction back in time. In these regards, it is noteworthy to say that ESM research often focuses on how situations and emotions differ from moment to moment and on their dynamic interplay (Krämer et al., 2023).

This view could reinforce the general idea that more recent forecasts provide the highest accuracy (Liberman et al., 2002), as well as partially contrasting the results of the experiment conducted by Finkenauer et al. (2007), which suggested that for negative affect, prediction accuracy would be higher when closer to the event and vice versa for positive affect.

If our results were found to support Finkenauer et al. (2007)'s findings, it would imply that aggregating forecasts for positive emotions could be beneficial because it would combine earlier, more accurate predictions with later ones. However, for negative emotions, aggregating earlier, less accurate predictions would not present any advantage.

In contrast, the findings of our exploratory analysis provide support for the results of our experiment, showing a significant difference between using the prediction 6 and 3 hours before each time point for either positive or negative affect, with the 3 hours ahead prediction leading in accuracy. Since no advantage was found in using older predictions against more recent ones, it makes sense that aggregation does not provide a significant advantage compared to using the most recent prediction. It is important to note that the discrepancy found between the two studies could be attributable to substantial differences in the conceptualization of the different variables and in the structure of the experiments. In fact, in the experiment conducted by Finkenauer et al. (2007), participants were asked to predict how well they would do on their driving license exam, an event that carries real emotional weight, whereas our experiment focused on ordinary, day-to-day mood predictions. These fundamental differences in the concept measured (exam performance vs everyday feelings) and in the structure of the two experiments make it unsurprising that the two studies do not line up exactly. On another note, we propose the possibility that the aggregation method employed in the current study may not have been optimal for the domain of affective forecasting. While in other domains (e.g.predicting the number of marbles in a glass jar) within-person aggregation works by canceling out random errors (Vul & Pashler, 2008), in the case of affective forecasting, the error is more driven by biases such as impact bias, which are systematic and thus persist across predictions from the same individual. This idea aligns with the findings of Satopää et al. 2021 who proposed the Bias, Information, Noise (BIN) model to break down forecasting error into systematic biases, information quality, and random noise. They suggested that aggregation would be more effective when random noise is the predominant factor, which is not the case in affective forecasting, proposing a targeted debiasing training as an alternative to reduce error in domains where it is mainly driven by systematic biases (Satopää et al., 2021). Although the same mechanism of averaging out random error might still be at play when aggregating in our domain, since systematic biases are more prevalent, it might not be sufficient to effectively diminish error in predictions. This could be the main reason why a significant improvement in accuracy through aggregation was not observed in our experiment.

It is important to mention that even in the field of human forecasting, where aggregation has been used for years, a general consensus has not been reached yet on the best method for aggregating predictions (McAndrew et al., 2021), and this may vary from one domain to another based on its particular characteristics. We suggest that a different method of aggregation, for example, one giving different weights to predictions produced at different time points, could yield different, possibly significant results.

Strengths

The use of ESM is an important strength as it represented a significant tool to largely increase the ecological validity of our study. By having participants respond to surveys throughout their daily routines, this tool allowed us to obtain the needed information in a way that closely mirrors real life, capturing fluctuations in contexts, mood, and emotional dynamics (Fritz et al., 2024).

Limitations

One limitation of the present study is the relatively small sample size of 29 participants. This limits the efforts made to employ robust statistical analyses, potentially resulting in an increased risk of Type II errors. Future research will have to employ larger sample sizes to increase the power in order to detect more subtle effects and enhance the generalizability of the findings.

An additional factor that might be seen as a limitation of the current study is the duration of the interval between predictions. Specifically, the fact that it is the first time investigating aggregation in this context meant that we had no previous knowledge from the same domain on what time interval could be ideal for effective aggregation. Studies on within-person aggregation in human forecasting consistently show that a time delay between predictions is important to obtain sufficient variation and therefore to average out random error (Vul & Pashler, 2008). The use of ESM allowed us to record predictions in real time and within specific and standardized time windows. It should be noted that the optimal amount of time between predictions, especially for affective forecasting, is unknown, and it is possible that the aggregation of more or less widely spaced predictions might yield different results. For example, it might be possible that the interval that was used in our experiment (3 hours between predictions) might not be enough to ensure enough independence between predictions. In their leading paper on within-person aggregation, for example, Vul and Pashler (2008) used intervals of 3 weeks between judgments, while Fraundorf and Benjamin (2014) had participants engage in tasks of a duration of around 30 minutes between tasks. Especially since the nature of the judgments produced in these experiments is also radically different from the ones investigated in our study, we suggest that future research should investigate aggregation in the context of affective forecasting within different time intervals, for example, days or minutes. This could be a useful addition to the field to improve our understanding of whether aggregation can actually become a useful intervention in the field, as well as its boundary conditions.

Future directions

Although not statistically significant, these findings provide a useful initial insight into the application of aggregation principles in affective forecasting. Moreover, future research could investigate the concept of aggregation in the affective forecasting domain in different contexts.

One of these could be to investigate affective forecasting aggregation in a similar structure to the one used by Finkenauer et al.(2007) in their experiment, moving the focus onto focal events rather than sampling daily forecasts. For example, participants could be asked to record predictions regarding recurrent focal events that carry some degree of emotional weight, for example, university exams or project check-ins in a work group. These predictions could be aggregated and then compared to the closest prediction in time to the focal event. Carrying such an experiment could also be an idea to experiment with different time intervals for aggregation, as predictions spaced between days could be obtained.

Conclusion

As of the current experiment, we can conclude that no predictive advantage was found for the use of aggregation in the context of affective forecasting. These findings reinforce the idea that, in this specific context, people are better at predicting events that are closer in time, and the closer the event is, the more accurate the prediction will be on average. We suggest that aggregation, which acts on random error, might have been at play in our experiment. However, the effect of systematic biases (e.g. impact bias), the main determinants of affective forecasting error, might have prevailed, making the impact of aggregation seem irrelevant, or non-significant on the final accuracy. It is important to highlight the fact that this research represents a starting point connecting the concept of aggregation with the field of affective forecasting. Future research will have to try to approach this concept in novel ways, exploring different aggregation methods, time intervals, and types of predicted events.

References

Armstrong, J. S. (2001). *Principles of forecasting: A handbook for researchers and practitioners*. Springer. https://doi.org/10.1007/978-0-306-47630-3

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Brown, V. A. (2021). An introduction to linear mixed-effects modeling in R. *Advances in Methods and Practices in Psychological Science, 4*(1), Article 2515245920960351. https://doi.org/10.1177/2515245920960351

Buehler, R., & McFarland, C. (2001). Intensity Bias in Affective Forecasting: The Role of Temporal Focus. *Personality and Social Psychology Bulletin*, *27*(11), 1480–1493. https://doi.org/10.1177/01461672012711009

Codi, A., Luk, D., Braun, D., Cambeiro, J., Besiroglu, T., Chen, E., Cesaris, de, Bocchini, P., & McAndrew, T. (2022). Aggregating Human Judgment Probabilistic Predictions of Coronavirus Disease 2019 Transmission, Burden, and Preventive Measures. *Open Forum Infectious Diseases*, *9*(8). https://doi.org/10.1093/ofid/ofac354

Csikszentmihalyi, M., Larson, R., & Prescott, S. (1977). The ecology of adolescent activity and experience. *Journal of Youth and Adolescence*, *6*(3), 281–294. https://doi.org/10.1007/BF02138940

Csikszentmihalyi, M., & Larson, R. (1987). Validity and reliability of the Experience Sampling Method. *Journal of Nervous and Mental Disease, 175*(9), 526–536. https://doi.org/10.1097/00005053-198709000-00004

Dunn, E. W., Brackett, M. A., Ashton-James, C., Schneiderman, E., & Salovey, P. (2007). On emotionally intelligent time travel: Individual differences in affective forecasting ability. *Personality and Social Psychology Bulletin*, *33*(1), 85–93. https://doi.org/10.1177/0146167206294201

Ellis, E. M., Elwyn, G., Nelson, W. L., Scalia, P., Kobrin, S. C., & Ferrer, R. A. (2018). Interventions to Engage Affective Forecasting in Health-Related Decision Making: A Meta-Analysis. *Annals of behavioral medicine : a publication of the Society of Behavioral Medicine*, *52*(2), 157–174. https://doi.org/10.1093/abm/kax024

Finkenauer, C., Gallucci, M., van Dijk, W. W., & Pollmann, M. (2007). Investigating the role of time in affective forecasting: Temporal influences on forecasting accuracy. *Personality* *and Social Psychology Bulletin*, *33*(8), 1152–1166. https://doi.org/10.1177/0146167207303021

Fiechter, J. L., & Kornell, N. (2021b). How the wisdom of crowds, and of the crowd within, are affected by expertise. *Cognitive Research: Principles and Implications*, *6*(1), 1–7. https://doi.org/10.1186/s41235-021-00273-6

Fraundorf, S. H., & Benjamin, A. S. (2014). Knowing the crowd within: Metacognitive limits on combining multiple judgments. *Journal of Memory and Language*, *71*(1), 17–38. https://doi.org/10.1016/j.jml.2013.10.002

Fritz, J., Piccirillo, M. L., Cohen, Z. D., Frumkin, M., Kirtley, O., Moeller, J., Neubauer, A. B., Norris, L. A., Schuurman, N. K., Snippe, E., & Bringmann, L. F. (2024). So You Want to Do ESM? 10 Essential Topics for Implementing the Experience-Sampling Method. *Advances in Methods and Practices in Psychological Science*, *7*(3). https://doi.org/10.1177/25152459241267912

Gilbert, D. T., Pinel, E. C., Wilson, T. D., Blumberg, S. J., & Wheatley, T. P. (1998). Immune neglect: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology, 75*(3), 617–638. https://doi.org/10.1037/0022-3514.75.3.617

Gilbert, D. T., & Wilson, T. D. (2007). Prospection: Experiencing the future. *Science*, *317*(5843), 1351–1354. https://doi.org/10.1126/science.1144161

Halpern, J., & Arnold, R. M. (2008). Affective forecasting: an unrecognized challenge in making serious health decisions. *Journal of general internal medicine*, *23*(10), 1708–1712. https://doi.org/10.1007/s11606-008-0719-5

Hoerger, M., Quirk, S. W., & Weed, N. C. (2012). Development and validation of the Delaying Gratification Inventory. *Psychological Assessment, 24*(3), 725–738. https://doi.org/10.1037/a0026636

Hoerger, M., Quirk, S. W., Lucas, R. E., & Carr, T. H. (2010). Cognitive determinants of affective forecasting errors. *Judgment and Decision Making*, *5*(5), 365–373. https://doi.org/10.1017/s1930297500002163

JASP Team (2024). JASP (Version 0.18.3)[Computer software].

Krämer, M. D., Roos, Y., Schoedel, R., Wrzus, C., & Richter, D. (2023). *Social dynamics and affect: Investigating within-person associations in daily life using experience sampling and mobile sensing*. Center for Open Science. https://doi.org/10.31234/osf.io/jxznd

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Liberman, N., Sagristano, Mi. D., & Trope, Y. (2002). The effect of temporal distance on level of mental construal. *Journal of Experimental Social Psychology*, *38*(6), 523–534. https://doi.org/10.1016/S0022-1031(02)00535-8

Litvinova, A., Herzog, S. M., Kall, A. A., Pleskac, T. J., & Hertwig, R. (2019). *How the "wisdom of the inner crowd" can boost accuracy of confidence judgments*. Center for Open Science. https://doi.org/10.31234/osf.io/ngweh

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). performance: An R Package for Assessment, Comparison and Testing of Statistical Models. *Journal of Open Source Software*, *6*(60). https://doi.org/10.21105/joss.03139

McAndrew, T., Wattanachit, N., Gibson, G. C., & Reich, N. G. (2021). Aggregating predictions from experts: a review of statistical methods, experiments, and applications - PMC. *Wiley Interdisciplinary Reviews. Computational Statistics*, *13*(2). https://doi.org/10.1002/wics.1514

Mellers, B. A., McCoy, J. P., Lu, L., & Tetlock, P. E. (2023b). Human and algorithmic predictions in geopolitical forecasting: Quantifying uncertainty in hard-to-quantify domains. *Perspectives on Psychological Science*, *19*(5), 711–721. https://doi.org/10.1177/17456916231185339

Mestdagh, M., Verdonck, S., Piot, M., Niemeijer, K., Kilani, G., Tuerlinckx, F., Kuppens, P., & Dejonckheere, E. (2023). m-Path: An easy-to-use and highly tailorable platform for ecological momentary assessment and intervention in behavioral research and clinical practice. *Frontiers in digital health*, *5*, Article 1182175. https://doi.org/10.3389/fdgth.2023.1182175

Pollai, M., Hoelzl, E., & Possas, F. (2009). Consumption-related emotions over time: Fit between prediction and experience. *Marketing Letters*, *21*(4), 397–411. https://doi.org/10.1007/s11002-009-9090-5

Posit team (2025). RStudio: Integrated Development Environment for R. Posit Software, PBC, Boston, MA. URL http://www.posit.co/.

Satopää, V., Salikhov, M., Tetlock, P., & Mellers, B. (2021). Bias, information, noise: The BIN model of forecasting. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3540864

Takano, K., & Ehring, T. (2024). Affective forecasting as an adaptive learning process. *Emotion*, *24*(3), 795–807. https://doi.org/10.1037/emo0001303

Ubel, P. A., Loewenstein, G., Schwarz, N., & Smith, D. (2005). Misimagining the unimaginable: The disability paradox and health care decision making. *Health Psychology, 24*(4, Suppl), S57–S62. https://doi.org/10.1037/0278-6133.24.4.S57

van Dolder, D., & van den Assem, M. J. (2017). The wisdom of the inner crowd in three large natural experiments. *Nature Human Behaviour*, *2*(1), 21–26. https://doi.org/10.1038/s41562-017-0247-6

Vul, E., & Pashler, H. (2008). Measuring the Crowd Within. *Psychological Science*, *19*(7), 645–647. https://doi.org/10.1111/j.1467-9280.2008.02136.x

Walsh, E., & Ayton, P. (2009). My imagination versus your feelings: Can personal affective forecasts be improved by knowing other peoples' emotions? *Journal of Experimental Psychology: Applied*, *15*(4), 351–360. https://doi.org/10.1037/a0017984

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43). https://doi.org/10.21105/joss.01686

Wilson, T. D., Wheatley, T., Meyers, J. M., Gilbert, D. T., & Axsom, D. (2000). Focalism: A source of durability bias in affective forecasting. *Journal of Personality and Social Psychology, 78*(5), 821–836. https://doi.org/10.1037/0022-3514.78.5.821

Wilson, T. D., Meyers, J., & Gilbert, D. T. (2001). Lessons from the Past: Do People Learn from Experience that Emotional Reactions Are Short-Lived? *Personality and Social Psychology Bulletin*, *27*(12), 1648–1661. https://doi.org/10.1177/01461672012712008

Wilson, T. D., & Gilbert, D. T. (2003). Affective forecasting. In *Advances in Experimental Social Psychology* (pp. 345–411). Elsevier. https://doi.org/10.1016/s0065-2601(03)01006-2

Appendix A

Code Absolute errors and GLMM

```
library(tidyverse)
library(lmerTest)
library(performance)

dat_prediction <- read.csv("results/full_predictions.csv")
dat_prediction_with_errors <- dat_prediction %>%
  filter(!is.na(PA_now_point) & !is.na(PA_aggregated_point_match)) %>%
```

```
  mutate(
    # --- Absolute Errors for Positive Affect (PA) ---
    # Error for "Most Recent" (1-step ahead matched) PA prediction
    Abs_Error_PA_MostRecent = abs(PA_now_point - PA_one_step_point_match),
    # Error for "Aggregated" PA prediction
    Abs_Error_PA_Aggregated = abs(PA_now_point - PA_aggregated_point_match),
    # --- Absolute Errors for Negative Affect (NA) ---
    # Error for "Most Recent" (1-step ahead matched) NA prediction
    Abs_Error_NA_MostRecent = abs(NA_now_point - NA_one_step_point_match),
    # Error for "Aggregated" NA prediction
    Abs_Error_NA_Aggregated = abs(NA_now_point - NA_aggregated_point_match)
  ) %>% select(ParticipantID:time_n, Abs_Error_PA_MostRecent:Abs_Error_NA_Aggregated)
dat_prediction_with_errors_long <- dat_prediction_with_errors %>%
  pivot_longer(
    cols = -c(ParticipantID:time_n),
    names_to = c("affect", "type"),
    names_pattern = "Abs_Error_(PA|NA)_(MostRecent|Aggregated)",
    values_to = "error"
  ) %>%
  pivot_wider(names_from = affect, values_from = "error",names_prefix = "error_")
write.csv(dat_prediction_with_errors_long,"results/full_glmm.csv")
mod_na <-  lmer(error_NA ~ type  + (1 | ParticipantID),
                          data = dat_prediction_with_errors_long)
```

```
summary(mod_na)
check_model(mod_na, panel = T,check = c("reqq","pp_check","qq","homogeneity"))
mod_pa <-  lmer(error_PA ~ type  + (1 | ParticipantID),
             data = dat_prediction_with_errors_long)
summary(mod_pa)
check_model(mod_pa, panel = T,check = c("reqq","pp_check","qq","homogeneity"))
# GLM
library(lme4)
mod_na_glm <-  glmer(error_NA + 1 ~ type  + (1 | ParticipantID),
             data = dat_prediction_with_errors_long,
             family = Gamma(link="log"))
summary(mod_na_glm)
check_model(mod_na_glm, panel = T,check = c("reqq","pp_check","qq","homogeneity"))
mod_pa_glm <-  glmer(error_PA + 1 ~ type  + (1 | ParticipantID),
              data = dat_prediction_with_errors_long,
              family = Gamma(link="log"))
summary(mod_pa_glm)
check_model(mod_pa_glm, panel = T,check = c("reqq","pp_check","qq","homogeneity"))
```

Code Compliance rates

```
library(dplyr)

expected_responses_total <- 5 * 14 # 70 questionnaires expected per participant
```

```
individual_compliance_rates_df <- full_predictions %>% # Use 'data_filtered' if that's your
refined dataset
  group_by(ParticipantID) %>%
  summarise(
    ResponsesSubmitted = sum(!is.na(PA_now_point)),
    ComplianceRate = (ResponsesSubmitted / expected_responses_total) * 100
  ) %>%
  ungroup()

average_compliance_rate <- mean(individual_compliance_rates_df$ComplianceRate)

min_compliance <- min(individual_compliance_rates_df$ComplianceRate)
max_compliance <- max(individual_compliance_rates_df$ComplianceRate)
sd_compliance <- sd(individual_compliance_rates_df$ComplianceRate)
median_compliance <- median(individual_compliance_rates_df$ComplianceRate)

cat("Individual Compliance Rates:\n")
print(individual_compliance_rates_df)

cat("\nOverall Average Compliance Rate:\n")
cat(paste0(round(average_compliance_rate, 2), "%\n"))
```

```
cat("\nDescriptive Statistics for Compliance Rates:\n")
cat(paste0("Min: ", round(min_compliance, 2), "%\n"))
cat(paste0("Max: ", round(max_compliance, 2), "%\n"))
cat(paste0("SD: ", round(sd_compliance, 2), "%\n"))
cat(paste0("Median: ", round(median_compliance, 2), "%\n"))

cat("\nFull Summary of Compliance Rates:\n")
print(summary(individual_compliance_rates_df$ComplianceRate))

cat("\nDetailed Descriptive Table:\n")
individual_compliance_rates_df %>%
  summarise(
    N_Participants = n(),
    Min_Compliance = min(ComplianceRate),
    Max_Compliance = max(ComplianceRate),
    Mean_Compliance = mean(ComplianceRate),
    Median_Compliance = median(ComplianceRate),
    SD_Compliance = sd(ComplianceRate)
  ) %>%
  print()
```

Code exploratory analysis

```
library(tidyverse)
```

```
dat_prediction <- read.csv("results/full_predictions.csv")
dat_prediction_with_errors <- dat_prediction %>%
  filter(!is.na(PA_now_point) & !is.na(PA_aggregated_point_match)) %>%
  mutate(
    # --- Absolute Errors for Positive Affect (PA) ---
    # Error for "Most Recent" (1-step ahead matched) PA prediction
    Abs_Error_PA_MostRecent = abs(PA_now_point - PA_one_step_point_match),
    # Error for "Oldest" PA prediction
    Abs_Error_PA_Aggregated = abs(PA_now_point - PA_aggregated_point_match),
    # --- Absolute Errors for Negative Affect (NA) ---
    # Error for "Most Recent" (1-step ahead matched) NA prediction
    Abs_Error_NA_MostRecent = abs(NA_now_point - NA_one_step_point_match),
    # Error for "Oldest" NA prediction
    Abs_Error_NA_Aggregated = abs(NA_now_point - NA_aggregated_point_match)
  ) %>% select(ParticipantID:time_n, Abs_Error_PA_MostRecent:Abs_Error_NA_Aggregated)
dat_prediction_with_errors_long <- dat_prediction_with_errors %>%
  pivot_longer(
    cols = -c(ParticipantID:time_n),
    names_to = c("affect", "type"),
    names_pattern = "Abs_Error_(PA|NA)_(MostRecent|Aggregated)",
    values_to = "error"
  ) %>%
  pivot_wider(names_from = affect, values_from = "error",names_prefix = "error_")
```

```
write.csv(dat_prediction_with_errors_long,"results/full_glmm.csv")
library(lmerTest)
library(performance)
mod_na <-  lmer(error_NA ~ type  + (1 | ParticipantID),
                         data = dat_prediction_with_errors_long)
summary(mod_na)
check_model(mod_na, panel = T,check = c("reqq","pp_check","qq","homogeneity"))
mod_pa <-  lmer(error_PA ~ type  + (1 | ParticipantID),
               data = dat_prediction_with_errors_long)
summary(mod_pa)
check_model(mod_pa, panel = T,check = c("reqq","pp_check","qq","homogeneity"))
# GLM
library(lme4)
mod_na_glm <-  glmer(error_NA + 1 ~ type  + (1 | ParticipantID),
               data = dat_prediction_with_errors_long,
               family = Gamma(link="log"))
summary(mod_na_glm)
check_model(mod_na_glm, panel = T,check = c("reqq","pp_check","qq","homogeneity"))


mod_pa_glm <-  glmer(error_PA + 1 ~ type  + (1 | ParticipantID),
                data = dat_prediction_with_errors_long,
                family = Gamma(link="log"))
```

```
summary(mod_pa_glm)

check_model(mod_pa_glm, panel = T,check = c("reqq","pp_check","qq","homogeneity"))
```

Appendix B

Figure 1

Assumption checks LMM Positive affect

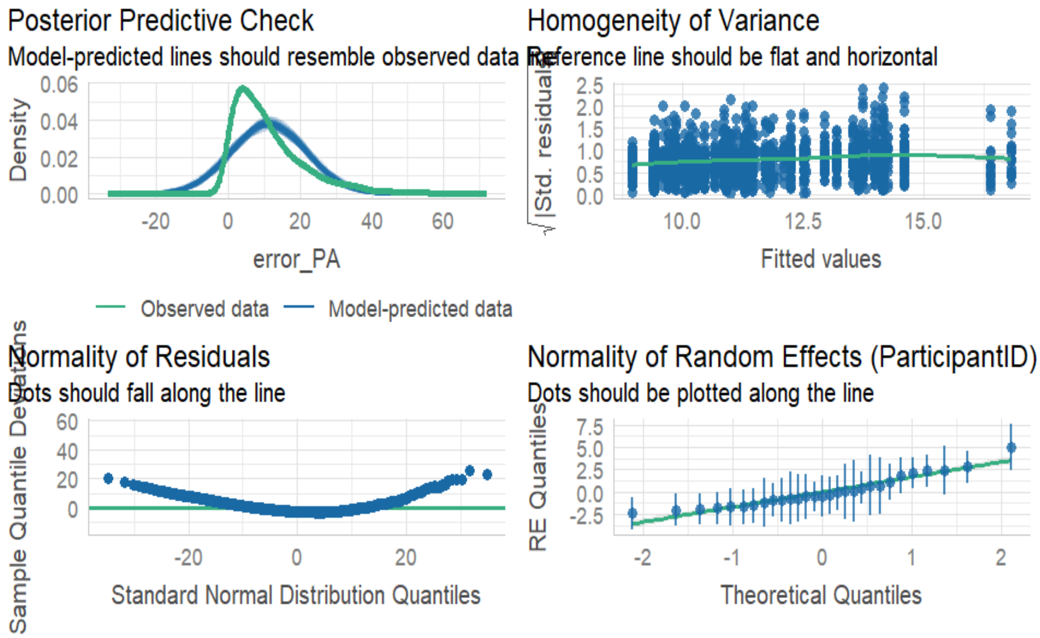


Figure 2

Assumption checks LMM Negative Affect

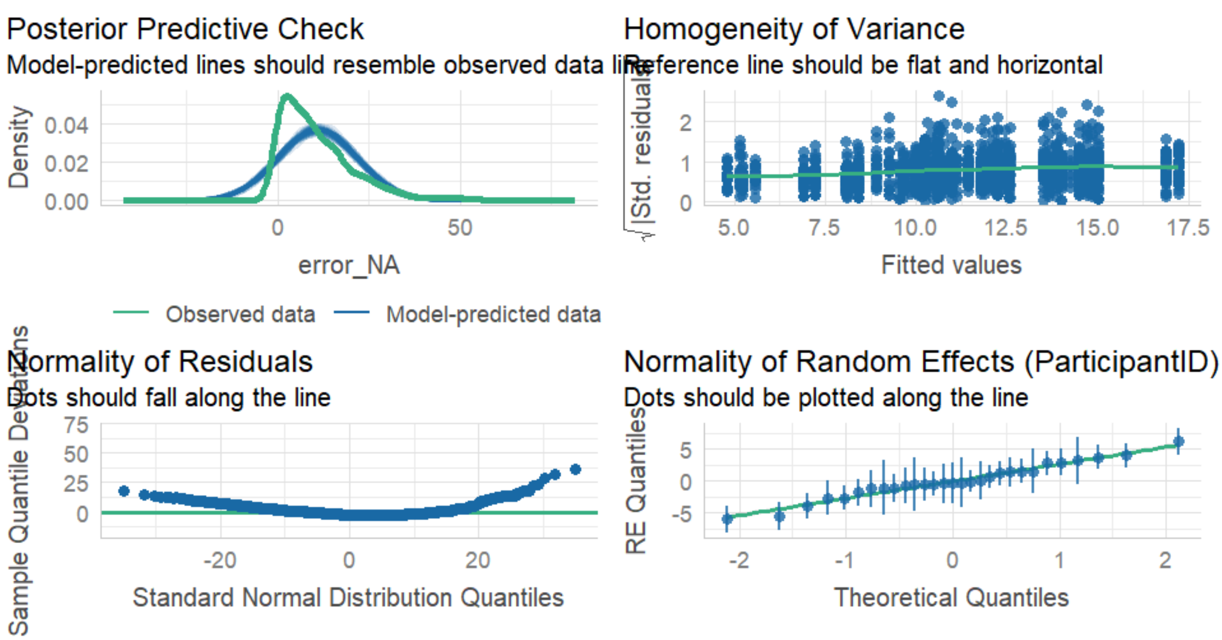


Figure 3

Assumptions checks GLMM Positive Affect

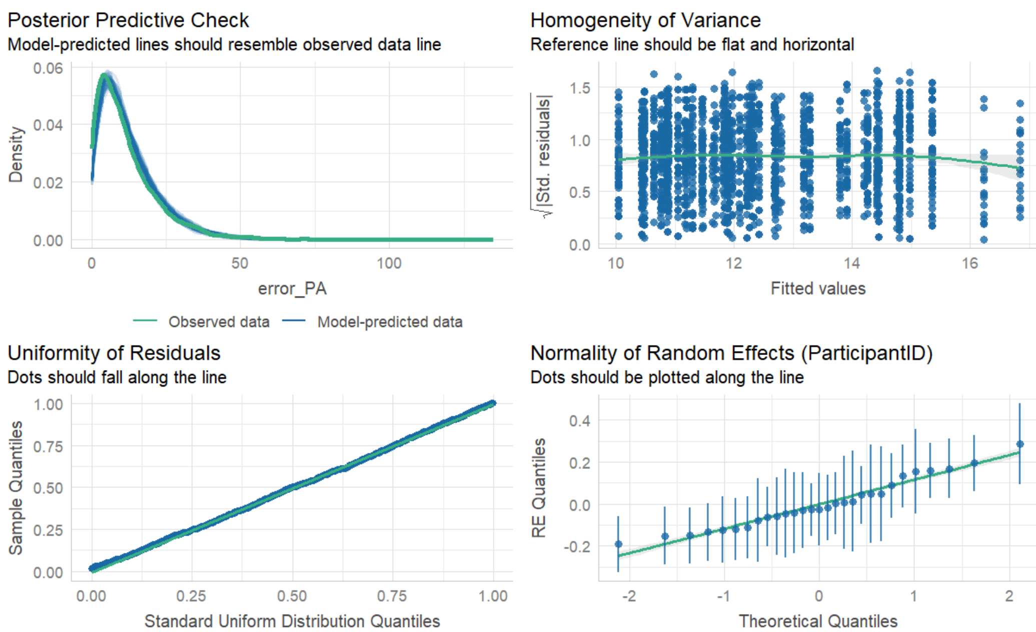
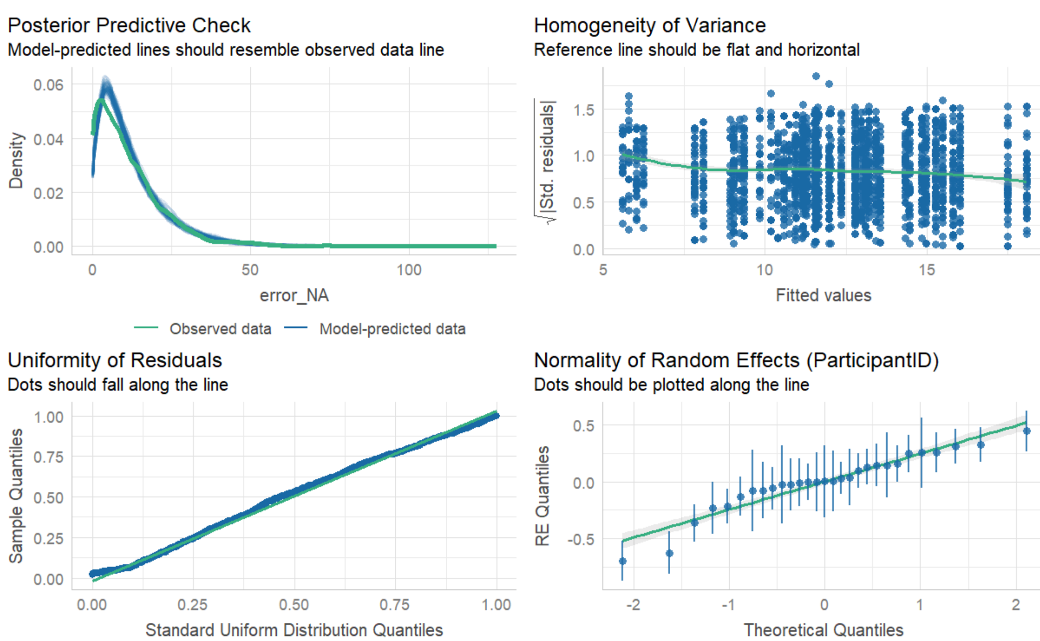


Figure 4

Assumptions checks GLMM negative affect



Appendix C

No content generated by AI technologies has been presented as my own work.

(1) I acknowledge the use of Gemini (https://gemini.google.com/app?hl=it) to generate materials for background research and self-study in the drafting of this assessment.

Input: is there a lot of studies in the field of affective forecasting investigating methods to improve accuracy?

Output: Yes, there are a lot of studies in the field of affective forecasting that investigate methods to improve accuracy. Researchers have explored various strategies to mitigate the common biases (like the impact bias and durability bias) that lead to inaccurate predictions of future feelings.

Some common methods and areas of investigation include:

- **Considering Alternatives/Debiasing Strategies:** Some research focuses on cognitive strategies that encourage individuals to consider a wider range of possibilities or contradictory information, similar to debiasing techniques in other judgment domains.
- **Averaging Multiple Forecasts (from the same person):** As discussed in the "Measuring the Crowd Within" paper, one surprising method being investigated is having individuals make multiple predictions for the same event and then averaging those predictions. The idea is that even within a single person, different "samples" from their internal probabilistic representation can reduce noise and improve accuracy.

- **Temporal Distance Manipulation:** Studies examine how varying the temporal distance to the event or the time between forecasts affects accuracy, with the aim of understanding and potentially leveraging these temporal dynamics.
- **Reducing Focalism:** Researchers explore ways to help people think more broadly about their future experiences, beyond just the focal event itself, to account for other life events and the process of adaptation that might influence their emotions.
- **Experience-Based Learning:** Investigating whether and how past experiences can lead to more accurate future forecasts, and what factors might hinder this learning process.
- **Using different modes of elicitation:** Exploring if different ways of asking people to predict their emotions (e.g., imagining vs. recalling similar past events) can yield more accurate results.

Use: This output was used to structure my research for papers regarding interventions to improve accuracy in affective forecasting

(2) I acknowledge the use of Gemini (https://gemini.google.com/app?hl=it) to get advice regarding statistical analysis on R

Input: “how do i perform a lmm to test for a difference between the error for Oldest vs MostRecent?

Output: To perform a Linear Mixed Model (LMM) to test for a difference between the "Oldest" and "MostRecent" error types, you'll need to specify your model carefully. Based on the R code *you provided, you are already using the* `lmerTest` *package, which is suitable for this purpose as it provides p-values for fixed effects.*

Here's how you can perform the LMM, focusing on comparing "Oldest" and "MostRecent" error types, and the interpretation:

1. Ensure 'MostRecent' is the Reference Level

For a direct comparison between "Oldest" and "MostRecent", it's helpful if "MostRecent" is the reference level for your `type` *factor. You can explicitly set the factor levels:*

R

```
None
dat_prediction_with_errors_long$type <-
factor(dat_prediction_with_errors_long$type, levels =
c("MostRecent", "Oldest", "Aggregated"))
```

In your provided code, you already have this line: `mutate(type = factor(type, levels = c("MostRecent", "Oldest", "Aggregated")))`*, so "MostRecent" is already set as the reference level.*

2. Perform the LMMs

You've already set up the LMMs for `error_NA` and `error_PA` in your code. The structure is correct for testing differences between `type` levels, with `ParticipantID` as a random intercept to account for repeated measures within participants.

For Negative Affect (NA):

R

```
None
mod_na <-  lmer(error_NA ~ type  + (1 | ParticipantID),

                          data =

dat_prediction_with_errors_long)

summary(mod_na)
```

For Positive Affect (PA):

R

```
None
mod_pa <-  lmer(error_PA ~ type  + (1 | ParticipantID),

                data = dat_prediction_with_errors_long)

summary(mod_pa)
```

3. Interpret the Results

After running `summary(mod_na)` or `summary(mod_pa)`, look at the "Fixed Effects" section of the output.

- ***Intercept**: This will represent the estimated mean error for the "MostRecent" `type` (because "MostRecent" is the reference level).*
- **`typeOldest`**: *This is the coefficient you are interested in.*
  - *The **Estimate** for `typeOldest` will represent the difference in mean error between the "Oldest" type and the "MostRecent" type.*
  - *The **p-value** associated with `typeOldest` will tell you if this difference is statistically significant. A small p-value (typically < 0.05) would suggest a significant difference between the "Oldest" and "MostRecent" error.*

Example Interpretation (Hypothetical):

If `summary(mod_na)` output for fixed effects showed:

```
None
Fixed Effects:

            Estimate Std. Error t value

(Intercept)   0.50       0.02   25.00

typeOldest    0.10       0.03    3.33  p-value = 0.001
```

```
typeAggregated -0.05       0.02   -2.50  p-value = 0.01
```

- *The average* `error_NA` *for "MostRecent" predictions is 0.50.*
- *The* `error_NA` *for "Oldest" predictions is, on average, 0.10 units higher than "MostRecent" predictions.*
- *The p-value of 0.001 for* `typeOldest` *indicates that this difference of 0.10 is statistically significant.*

This approach directly addresses your question of testing for a difference between the error for "Oldest" vs. "MostRecent" by setting "MostRecent" as the reference and examining the coefficient for "typeOldest".

Use: This output was used to help me perform the exploratory analysis through RStudio

(3) I acknowledge the use of Gemini (https://gemini.google.com/app?hl=it) to summarize and gather information from uploaded papers.

Input: “What does this paper say about affective forecasting accuracy in relation to time and positive and negative affect?”

Output: The paper "Investigating the Role of Time in Affective Forecasting: Temporal Influences on Forecasting Accuracy" by Finkenauer, Gallucci, van Dijk, and Pollmann (2007) explores how the time course of affective reactions and temporal distance to an event influence the accuracy of affective forecasting for both positive and negative affect.

Key findings regarding affective forecasting accuracy in relation to time and positive/negative affect include:

- **General Inaccuracy (Impact Bias):** People generally overestimate the intensity and duration of their emotional reactions to future events, a phenomenon known as the "impact bias". This bias is more pronounced for negative events than for positive ones. For example, participants in the study grossly overestimated the intensity and duration of their disappointment after failing a driving exam, but only slightly overestimated their happiness after passing.
- **Time Course of Affective Reactions:**
    - **Linear Changes:** People accurately predict that their affective reactions (both positive and negative) will decrease in intensity over time.
    - **Quadratic Changes (Speed of Decrease):** However, people tend to underestimate the *speed* with which these affective reactions will decrease. Their forecasts often reflect a linear decline rather than the more rapid, quadratic decline observed in actual experiences, suggesting their theories about affect progression are inaccurate regarding the *speed* of change.
- **Influence of Temporal Distance to the Focal Event:**
    - **Positive Affect:** Forecasting accuracy for positive affect is greater when the event is psychologically distant. As the event draws nearer, the accuracy of predicting positive affect decreases, and people are more likely to overestimate its intensity and duration.

- **Negative Affect:** Conversely, forecasting accuracy for negative affect is lower when the event is distant, but it *increases* as the event draws closer. When the event is close, people more accurately predict that their negative affect will fade, and importantly, that it will fade quickly.
- In summary, while people understand that their emotions will diminish over time, they often fail to account for the rapid initial decline. The temporal distance to an event also plays a crucial role, with people being more accurate at forecasting positive affect when the event is distant and more accurate at forecasting negative affect when the event is near

Use: The output was used to clarify my understanding of this paper