nature portfolio

https://doi.org/10.1038/s41562-022-01390-y

Supplementary information

Amplification in the evaluation of multiple emotional expressions over time

Amplification in the Evaluation of Multiple Emotional Expressions Over Time
(Supplementary Information)

1. Pre-Test: Evaluation of Amplification in a Single Face Trials

The goal of this pre-test was to examine whether amplification can be found in the evaluation of a single face within the Radboud face sample [1]. Similar analysis have been made with the NimStim stimuli in a previous paper on estimation of crowds' emotions[2]. Because we hypothesized that amplification is caused by processes relating to ensemble coding of sequences, we wanted to rule out the possibility that amplification occurred in the evaluation of a single face.

Method

Participants. Similar to all other studies we recruited 100 participants on prolific in exchange for \$2.3. Informed consent was obtained by participants. Two participants were removed from the analysis for providing average extreme ratings of less than 10 or more than 40, leaving the final sample at N = 98 (men: 39, women: 58, other: 1; Age: M= 25.85, SD = 9.19).

Procedure. The procedure was identical to that of Studies 1 with one difference, which is that participants only saw a single face in each trial.

Measures. The measures were identical to that used in Study 1.

Results and Discussion

We first tested the correlation between participants' estimation of the face intensity and the actual face intensity. Results suggested that the correlation was *r = .92* [.91, .92]. To measure amplification in the estimation of emotion on a single face, we conducted a mixed model analysis of repeated measures, comparing between the actual emotion expressed on a face, and the estimated emotion. As each participant was exposed to four types of face identities, we added two random variables, one for face-identity and one for participants' id. Results suggested that there was no significant difference between participants' estimation of the emotion on the face and the actual emotion on the face (*b* = 2.87, *t*(9,660) = -0.38, *p* = .70, R^2 = .x, 95% Confidence Intervals = [-1.35, .90]).

Next, we set out to test whether amplification was stronger for positive versus negative emotions. Similar to Studies 1-d, we created a difference score between participants' estimation of the estimated emotion on the face and the actual emotion, such that positive numbers indicated amplification. We then conducted a mixed model analysis, with valence predicting the degree of difference between estimated and actual mean emotions. As in our previous analysis, we used by-face-identity and by-participant random variables. Results suggested that there was no significant difference between estimation of positive and negative faces (*b* = 0.21, *t*(4808) = 0.67, *p* = .49, R^2 =

.x, 95% Confidence Intervals = [-0.39, 0.82]). Results from our analysis support the claim that amplification does not occur at the individual face level.

Finally, we set out to test whether amplification was stronger for male versus female faces. We conducted a similar analysis to that of the valence. Results suggested that there was no significant difference between estimation of positive and negative faces (*b* = -.33, *t*(4808) = -.83, *p* = .43, R^2 = .x, 95% Confidence Intervals = [-1.13, 0.48]).

2. Studies 1-4: Establishing Amplification in the Evaluation of Sequence of Facial Expressions

Amplification Based on Sequence Length

After demonstrating that amplification effect increases with sequence length, we investigated at which length participants start to overestimate the emotionality of a sequence. First, we calculated the mean estimation difference for each sequence length for Studies 1 – 4 by subtracting participants' estimation of the mean sequence emotion and the actual mean sequence emotion (Supplementary Table 1, Supplementary Table 3, Supplementary Table 5, Supplementary Table 7 and Supplementary Figures 1 – 4). Second, to statistically test amplification effect for certain sequence length, we ran the same model as specified by H1 in the main manuscript, dividing the sequence length into three categories: sequences with 4 or less faces, 4 – 6 faces, and more than 6 faces (Supplementary Table 2, Supplementary Table 4, Supplementary Table 6, Supplementary Table 8). This split was determined by visual inspection of the mean estimation differences; that is, 4 - 6 was the sequence length where amplification starts to occur. This statistical test was a mixed model analysis of repeated measures, comparing the actual mean emotion expressed in each set with participants' estimated mean emotion. We added a by-participant and by-face-identity random intercepts.

Study 1: Establishing effect

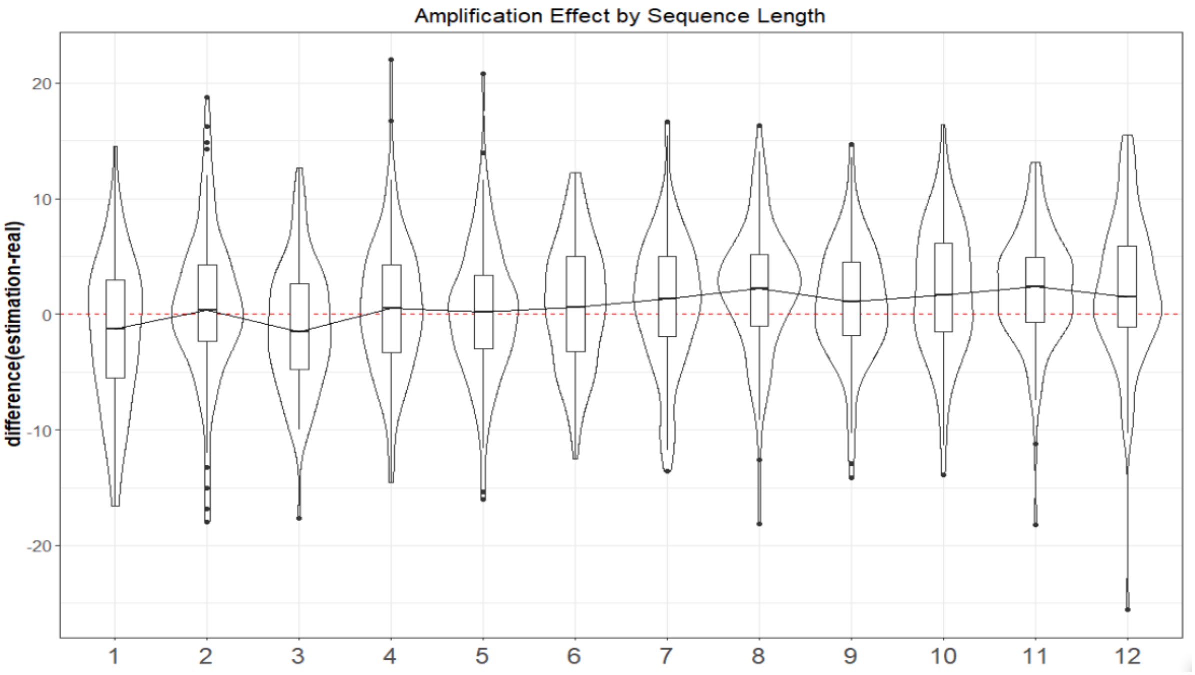
Supplementary Table 1. Overview of estimation differences for certain sequence length.

Sequence Length	Average Estimation Difference	SD Estimation Difference	Number of Trials
1 Face	-2.07	12.74	400
2 Faces	0.13	9.85	398
3 Faces	-1.04	9.93	406
4 Faces	0.29	9.68	406
5 Faces	-0.13	9.55	404
6 Faces	1.18	9.12	388
7 Faces	0.97	9.02	406

8 Faces	2.30	8.58	428
9 Faces	1.18	8.88	427
10 Faces	1.62	8.90	420
11 Faces	1.99	9.00	416
12 Faces	2.48	9.12	400

Supplementary Table 2. Overview of estimation differences for certain sequence length. The table shows the results of the linear mixed models as specified by H1 in the main manuscript, dividing the sequence length into three categories: sequences with 4 or less faces, 4 – 6 faces, and more than 6 faces. A one-sided t-test is used for statistical testing of the coefficients. Due to multiple comparisons the α is Bonferroni adjusted such that $p_{adjusted}$ = .017.

Sequence Length	**b [ci], (se)**	**t (df)**	**p**	**R^2**
≤ 4 Faces	-0.67 [-1.42 - 0.07], (0.38)	-1.76 (3214.99)	0.08	.0028
4 - 6 Faces	0.43 [-0.85 - 0.29], (0.33)	1.28 (2390.86)	.19	.0011
> 6 Faces	1.75 [1.35 - 2.16], (0.21)	8.52 (4989.25)	.001***	.014



Supplementary Figure 1. Violin plots showing the average estimation difference for each sequence length. The box plots display the median, first, and third quartiles. The whiskers extend to the most extreme value less than 1.5 times the interquartile range beyond the quartile. Outliers, which are data points outside the 1.5 interquartile range, are plotted as individual dots. The total number of participants is n = 93. A line connects the medians of each sequence length indicating an increase in sequence length is associated with an increase in amplification. Values above the red-dotted line are amplified, suggesting that amplification starts occurring between 4 – 6 faces per sequence.

Study 2: Replication with a new morph set

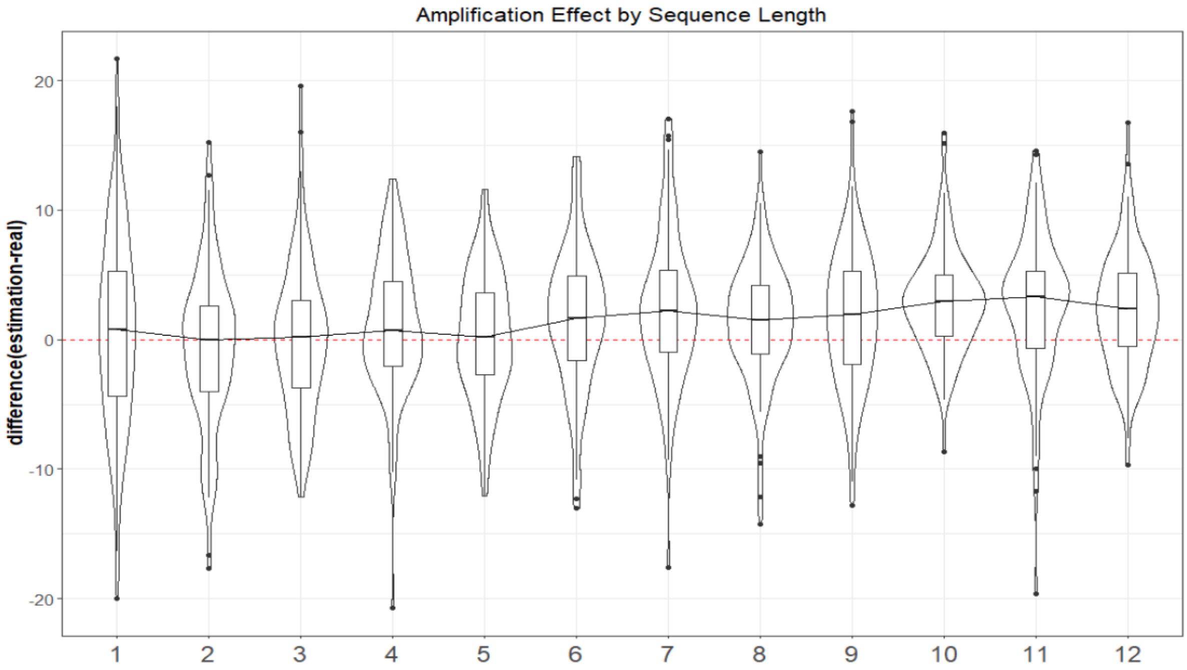
Supplementary Table 3. Overview of estimation differences for certain sequence length.

Sequence Length	Average Estimation Difference	SD Estimation Difference	Number of Trials
1 Face	0.51	11.59	361
2 Faces	-0.70	9.66	365
3 Faces	-0.35	9.61	388
4 Faces	1.10	8.59	383
5 Faces	0.52	8.50	372
6 Faces	1.54	8.52	361
7 Faces	1.97	8.70	389
8 Faces	1.44	7.59	371
9 Faces	1.67	8.15	356
10 Faces	3.13	7.19	389
11 Faces	2.99	7.53	385
12 Faces	2.48	7.17	372

Supplementary Table 4. Overview of estimation differences for certain sequence length. The table shows the results of the linear mixed models as specified by H1 in the main manuscript, dividing the sequence length into three categories: sequences with 4 or less faces, 4 – 6 faces, and more than 6 faces. A one-sided t-test is used for statistical testing of the coefficients. Due to multiple comparisons the α is Bonferroni adjusted such that $p_{adjusted}$ = .017.

Sequence Length	b [ci], (se)	t (df)	p	R²
≤ 4 Faces	0.14 [-0.56 - 0.85], (0.36)	0.40 (2984.94)	0.68	.00072
4 - 6 Faces	1.05 [0.41 – 1.69], (0.33)	3.19 (2230)	.001***	.0045
> 6 Faces	2.29 [1.91 - 2.67], (0.19)	11.81	.001***	0.029

		(4522)		



Supplementary Figure 2. Violin plots showing the average estimation difference for each sequence length. The box plots display the median, first, and third quartiles. The whiskers extend to the most extreme value less than 1.5 times the interquartile range beyond the quartile. Outliers, which are data points outside the 1.5 interquartile range, are plotted as individual dots. The total number of participants is n = 94. A line connects the medians of each sequence length indicating an increase in sequence length is associated with an increase in amplification. Values above the red-dotted line are amplified, suggesting that amplification starts occurring between 4 – 6 faces per sequence.

Study 3: Scale starts on right side

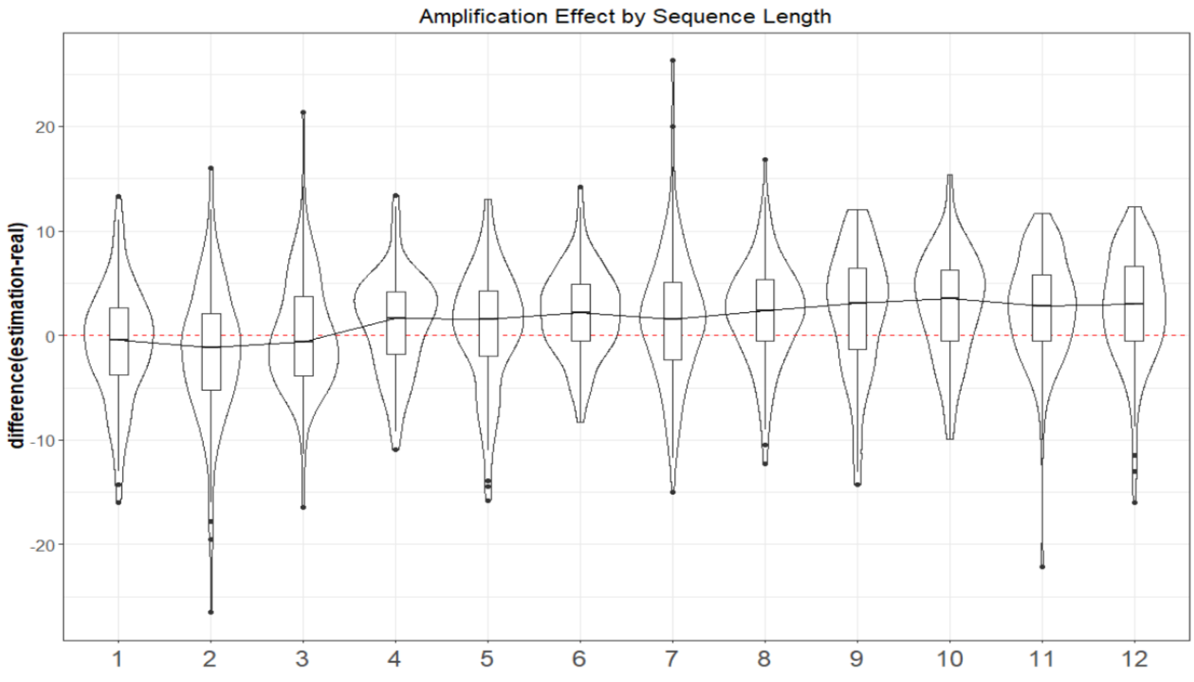
Supplementary Table 5. Overview of estimation differences for certain sequence length.

Sequence Length	**Average Estimation Difference**	**SD Estimation Difference**	**Number of Trials**
1 Face	-0.60	10.16	443
2 Faces	-1.33	9.24	396
3 Faces	-0.12	9.72	424
4 Faces	0.87	8.97	411
5 Faces	0.88	9.41	411
6 Faces	2.11	8.42	476

7 Faces	1.29	9.16	404
8 Faces	2.03	8.68	407
9 Faces	2.67	8.90	431
10 Faces	2.81	8.60	417
11 Faces	2.24	8.49	427
12 Faces	2.84	8.20	409

Supplementary Table 6. Overview of estimation differences for certain sequence length. The table shows the results of the linear mixed models as specified by H1 in the main manuscript, dividing the sequence length into three categories: sequences with 4 or less faces, 4 – 6 faces, and more than 6 faces. A one-sided t-test is used for statistical testing of the coefficients. Due to multiple comparisons the α is Bonferroni adjusted such that $p_{adjusted}$ = .017.

Sequence Length	b [ci], (se)	t (df)	p	R²
≤ 4 Faces	-0.28 [-1.03 - 0.45], (0.37)	-0.76 (3342.94)	0.44	.0022
4 - 6 Faces	1.32 [0.69 – 1.95], (0.32)	4.13 (2591)	.001***	.011
> 6 Faces	2.32 [1.92 - 2.71], (0.20)	11.49 (4985)	.001***	.028



Supplementary Figure 3. Violin plots showing the average estimation difference for each sequence length. The box plots display the median, first, and third quartiles. The whiskers extend to the most extreme value less than 1.5 times the interquartile range beyond the quartile. Outliers, which are data points outside the 1.5 interquartile range, are plotted as individual dots. The total number of participants is n = 98. A line connects the medians of each sequence length indicating an increase in sequence length is associated with an increase in amplification. Values above the red-dotted line are amplified, suggesting that amplification starts occurring between 4 – 6 faces per sequence.

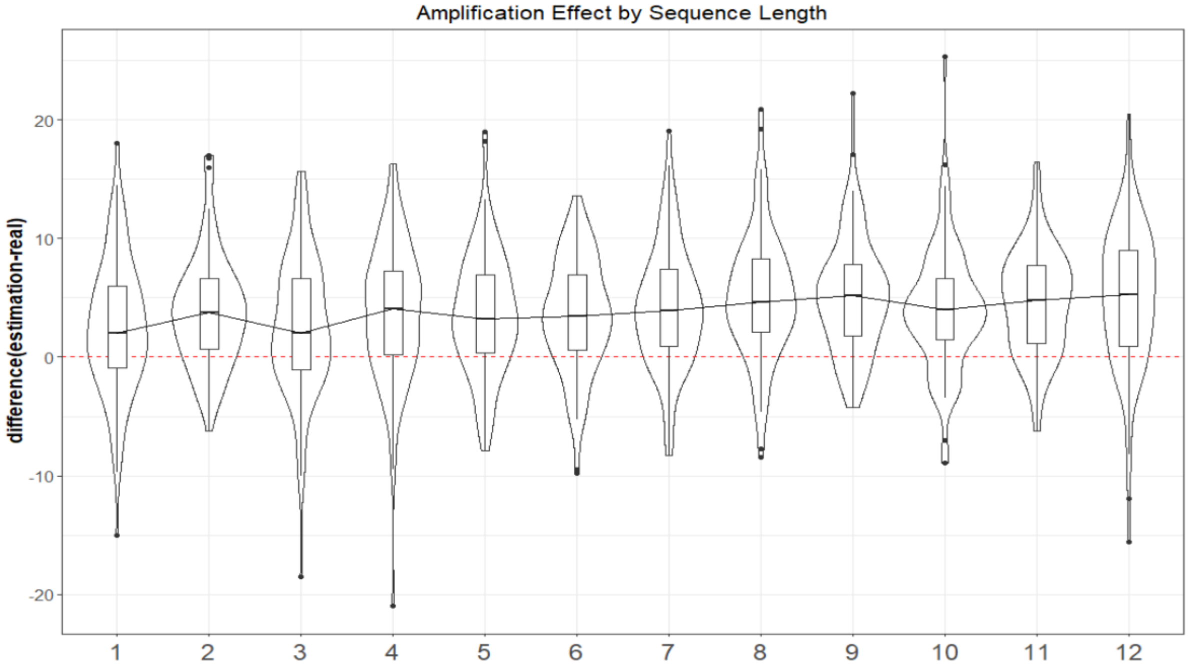
Study 4: Scale starts with strong intensity

Supplementary Table 7. Overview of estimation differences for certain sequence length.

Sequence Length	**Average Estimation Difference**	**SD Estimation Difference**	**Number of Trials**
1 Face	2.01	10.81	420
2 Faces	4.11	9.83	421
3 Faces	2.46	9.80	377
4 Faces	3.69	8.86	389
5 Faces	4.05	9.16	398
6 Faces	3.51	8.68	374
7 Faces	4.29	8.99	400
8 Faces	4.65	8.98	388
9 Faces	5.01	8.22	413
10 Faces	4.13	8.73	352
11 Faces	4.59	8.27	425
12 Faces	4.94	8.87	387

Supplementary Table 8. Overview of estimation differences for certain sequence length. The table shows the results of the linear mixed models as specified by H1 in the main manuscript, dividing the sequence length into three categories: sequences with 4 or less faces, 4 – 6 faces, and more than 6 faces. A one-sided t-test is used for statistical testing of the coefficients. Due to multiple comparisons the α is Bonferroni adjusted such that $p_{adjusted}$ = .017.

Sequence Length	**b [ci], (se)**	**t (df)**	**p**	**R^2**
≤ 4 Faces	3.07 [2.32 – 3.82], (0.38)	8.03 (3212)	.001***	.019
4 - 6 Faces	3.75 [3.07 – 4.44], (0.35)	10.78 (2316.92)	.001***	.053
> 6 Faces	4.61 [4.20 – 5.02], (0.21)	21.92 (4724.97)	.001***	.095



Supplementary Figure 4. Violin plots showing the average estimation difference for each sequence length. The box plots display the median, first, and third quartiles. The whiskers extend to the most extreme value less than 1.5 times the interquartile range beyond the quartile. Outliers, which are data points outside the 1.5 interquartile range, are plotted as individual dots. The total number of participants is n = 92. A line connects the medians of each sequence length indicating an increase in sequence length is associated with an increase in amplification. Values above the red-dotted line are amplified, suggesting amplification occurs for all trials when the estimation scale begins with displaying intense emotions.

Amplification and Potential Moderators (Correlation with Scales)

We examined if the amplification effect was moderated by demographic variables such as gender and race or individual differences as assessed by a few scales: social interaction anxiety scale[3] , PANAS scale [4], Big-Five personality scale[5], the UCLA three-item loneliness scale[6]. In all of our analyses we created a difference score between participants' estimation of the mean crowd emotion and the actual mean crowd emotion, such that positive numbers indicated amplification. We then conducted a series of mixed model analyses using each potential moderator as a predictor, and the difference score as the outcome variable. As in our previous analysis, we included by-face-identity and by-participant random variables.

Demographics as Moderators for Amplification

Race. Participants were asked to indicate their race from 5 options: Black, Hispanic, White, Asian and other, which they could specific individually. Participants could select multiple options for this question. As both participants as well as displayed faces were white, we binarized the options into "white" and "other races" participants to increase the possibility of finding a moderator effect. The results displayed in Supplementary Table 10 shows the coefficient for being white compared to identifying oneself as any of the other races.

Supplementary Table 9. Overview of number of participants identifying with a certain race.

Study	***N***	**Black**	**Hispanic**	**White**	**Asian**	**Other**
1: Establishing effect	93	1	11	75	1	5
2: Replication with a new morph set	94	5	2	75	7	5
3: Scale starts on right side	98	0	4	85	2	7
4: Scale starts with strong intensity	92	2	13	68	3	6

Supplementary Table 10. The coefficients of race (binarized race as white and other races) predicting participants tendency to amplify (difference between actual intensity of sequence and participants estimation) using a linear mixed model with participants as random intercept. The coefficient shows the difference of being white compared to identifying themselves as any other race. A one-sided t-test is used for statistical testing of the coefficients.

Study	**b [ci], (se)**	**t (df)**	**p**	**R^2**
1: Establishing effect	0.70 [-1.17, 2.57], (0.96)	0.95 (91.03)	.47	.13
2: Replication with a new morph set	1.63 [0.030, 3.23], (0.81)	1.99 (92.01)	.048*	.12
3: Scale starts on right side	0.53 [-1.27, 2.34], (0.92)	0.57 (95.79)	.57	.099
4: Scale starts with strong intensity	0.083 [-1.16, 1.33], (0.63)	0.13 (89.66)	.89	.067

Gender. Participants were asked to indicate their gender from 3 options: male, female or prefer not to say. The results displayed in Supplementary Table 12 shows the coefficient for being female compared to identifying oneself as male or other.

Supplementary Table 11. Overview of number of participants identifying with a certain race.

Study	***N***	**Male**	**Female**	**Prefer not to say**
1: Establishing effect	93	38	45	0
2: Replication with a new morph set	94	35	55	4
3: Scale starts on right side	98	62	35	1
4: Scale starts with strong intensity	92	62	30	0

Supplementary Table 12. Model coefficients of gender predicting participants tendency to amplify (difference between actual intensity of sequence and participants estimation) using a linear mixed model with a random intercept for participants. The coefficient shows the difference of being "female" compared to identifying themselves as "male". A one-sided t-test is used for statistical testing of the coefficients.

Study	**b [ci], (se)**	**t (df)**	**p**	**R^2**
1: Establishing effect	-0.46 [-0.80, 2.13], (0.76)	-0.60 (91.04)	.55	.13
2: Replication with a new morph set	-0.65 [-1.97, 0.67], (0.67)	-0.96 (92.13)	.33	.12
3: Scale starts on right side	-1.33 [-2.59, -0.074], (0.64)	-2.07 (96.00)	.04*	.099
4: Scale starts with strong intensity	0.64 [-0.51, 1.80], (0.59)	1.08 (89.77)	.28	.067

Individual Differences Scales as Predictors for Amplification

Social Anxiety. We measured social anxiety using an abbreviated version of the social interaction anxiety[3] scale which includes the following items on a scale from (0 - Not at all to 4-Extremely)

1. *I have difficulty making eye-contact with others.*
2. *I find it difficult mixing comfortably with the people I work with.*
3. *I tense-up if I meet an acquaintance on the street.*
4. *I feel tense if I am alone with just one person.*

5. *I have difficulty talking with other people.*
6. *I find it difficult to disagree with another's point of view.*

Supplementary Table 13. Ns, means, and standard deviations of the Social Anxiety Questionnaire (SIAS-6) across studies 1-4.

Study	***n***	***M***	***SD***	**α [ci]**
1: Establishing effect	93	2.56	0.65	0.86 [0.89,0.92]
2: Replication with a new morph set	-	-	-	-
3: Scale starts on right side	98	1.83	0.64	0.90 [0.87,0.92]
4: Scale starts with strong intensity	92	1.81	0.64	0.87 [0.84,0.91]

We then used multilevel models to predict amplification (difference between estimation and actual mean of the sequence) by their social anxiety score. We again added a random intercept of participants' id to the model.

Supplementary Table 14. Model coefficients of social anxiety score predicting tendency to amplify (difference between actual intensity of sequence and participants estimation) using a linear mixed model with a random intercept for participants. The coefficient shows the correlation coefficient of social anxiety score. A one-sided t-test is used for statistical testing of the coefficients.

Study	**b [ci], (se)**	**t (df)**	**p**	**R^2**
1: Establishing effect	-0.38 [-1.48, 0.73], (0.56)	-0.66 (90.98)	.51	.13
2: Replication with a new morph set	-	-	-	.12
3: Scale starts on right side	-0.067 [-1.03, 0.89], (0.49)	-0.13 (90.03)	.89	.099
4: Scale starts with strong intensity	0.17 [-0.70, 1.06], (0.45)	0.39 (82.89)	.69	.067

Personality. We assessed participants personality using the ten item personality scale (TIPI, Gosling et al., 2003). Each of the five personality factors has two items in this scale. One of the items is always reverse coded. Participants rate themselves on how much these two items apply to themselves on a scale from 1- Strongly Disagree to 7-Strongly Agree. The score is the average of two items (after reversing one item).

Extraversion: Is being assessed by the items:

1. *Extraverted, enthusiastic*
2. *Reserved, quiet*

Supplementary Table 15. Ns, means, and standard deviations of the Big Five Personality Questionnaire Factor Extraversion across studies 1-4. As Cronbach's Alpha requires 3 or more items we show the between item correlation (*r*) instead.

Study	***n***	***M***	***SD***	***r* [ci]**
1: Establishing effect	93	3.61	1.36	-0.33 [-0.55, -0.14]
2: Replication with a new morph set	94	3.64	1.51	-0.53 [-0.66, -0.37]
3: Scale starts on right side	98	3.52	1.48	-0.34 [-0.50, -0.15]
4: Scale starts with strong intensity	92	3.55	1.41	-0.33 [-0.50, -0.14]

We then used multilevel models to predict amplification (difference between estimation and actual mean of the sequence) by their extraversion score. We again added a random intercept for participants.

Supplementary Table 16. The coefficients of the moderator (Big Five extraversion score) and participants tendency to amplify (difference between actual intensity of sequence and participants estimation) using a linear mixed model with a random intercept for participants. The coefficient shows the correlation coefficient of extraversion score. A one-sided t-test is used for statistical testing of the coefficients.

Study	**b [ci], (se)**	**t (df)**	**p**	**R^2**
1: Establishing effect	-0.086 [-0.63, 0.46], (0.28)	-0.31 (90.96)	.76	.13
2: Replication with a new morph set	-0.064 [-0.50, 0.37], (0.22)	-0.29 (92.40)	.77	.12
3: Scale starts on right side	-0.025 [-0.44, 0.39], (0.21)	-0.12 (95.86)	.90	.099

4: Scale starts with strong intensity	-0.15 [-0.54, 0.23], (0.23)	-0.78 (89.67)	.43	.067

Neuroticism: Is being assessed by the items:

1. *Anxious, easily upset.*
2. *Calm, emotionally stable.*

Supplementary Table 17. Ns, means, and standard deviations of the Big Five personality questionnaire factor neuroticism across Studies 1-4. As Cronbach's Alpha requires 3 or more items we show the between item correlation (*r*) instead.

Study	***n***	***M***	***SD***	***r* [ci]**
1: Establishing effect	93	3.72	1.47	-0.42 [-0.57, -0.23]
2: Replication with a new morph set	94	3.80	1.44	-0.61 [-0.72, -0.46]
3: Scale starts on right side	98	3.86	1.44	-0.46 [-0.60, -0.28]
4: Scale starts with strong intensity	92	3.78	1.25	-0.24 [-0.43, -0.045]

We then used multilevel models to predict amplification (difference between estimation and actual mean of the sequence) by their neuroticism score. We again added a random intercept for participants' id.

Supplementary Table 18. The coefficients of the moderator (Big Five neuroticism score) and participants tendency to amplify (difference between actual intensity of sequence and participants estimation) using a linear mixed model with a random intercept for participants. The coefficient shows the correlation coefficient of neuroticism score. A one-sided t-test is used for statistical testing of the coefficients.

Study	**b [ci], (se)**	**t (df)**	**p**	**R^2**
1: Establishing effect	-0.50 [-1.00, -0.0093], (0.24)	-2.00 (90.98)	.048*	.13
2: Replication with a new morph set	-0.33 [-0.78, 0.11], (0.24)	-1.45 (92.18)	.15	.12
3: Scale starts on right side	-0.36 [-0.78, 0.061], (0.22)	-1.67 (96.02)	.09	.099

4: Scale starts with strong intensity	0.23 [-0.19, 0.67], (0.22)	1.08 (89.74)	.28	.067

Agreeableness: Is being assessed by the items:

1. *Critical, quarrelsome.*
2. *Sympathetic, warm.*

Supplementary Table 19. Ns, means, and standard deviations of the Big Five Personality questionnaire factor agreeableness across Studies 1-4. As Cronbach's Alpha requires 3 or more items, we show the between item correlation (*r*) instead.

Study	***n***	***M***	***SD***	***r* [ci]**
1: Establishing effect	93	3.19	1.09	-0.42 [-0.57, -0.23]
2: Replication with a new morph set	94	3.29	1.14	-0.61 [-0.72, -0.46]
3: Scale starts on right side	98	3.40	0.95	-0.46 [-0.60, -0.28]
4: Scale starts with strong intensity	92	3.45	1.06	-0.24 [-0.43, -0.045]

We then used multilevel models to predict amplification (difference between estimation and actual mean of the sequence) by their agreeableness score. We again added a random intercept for participants' id.

Supplementary Table 20. The coefficients of the moderator (Big Five agreeableness score) and participants tendency to amplify (difference between actual intensity of sequence and participants estimation) using a linear mixed model with a random intercept for participants. The coefficient shows the correlation coefficient of agreeableness score. A one-sided t-test is used for statistical testing of the coefficients.

Study	**b [ci], (se)**	**t (df)**	**p**	**R^2**
1: Establishing effect	0.19 [-0.49, 0.88], (0.35)	0.56 (91.01)	.58	.13
2: Replication with a new morph set	0.22 [-0.34, 0.80], (0.29)	0.78 (91.99)	.44	.12
3: Scale starts on right side	0.41 [-0.23, 1.05], (0.32)	1.25 (95.97)	.21	.099

4: Scale starts with strong intensity	-0.39 [-0.90, 0.12], (0.26)	-1.49 (89.61)	.14	.067

Conscientiousness: Is being assessed by the items:

1. *Dependable, self-disciplined.*
2. *Disorganized, careless*

Supplementary Table 21. Ns, means, and standard deviations of the Big Five personality questionnaire factor Conscientiousness across Studies 1-4. As Cronbach's Alpha requires 3 or more items we show the between item correlation (*r*) instead.

Study	***n***	***M***	***SD***	***r* [ci]**
1: Establishing effect	93	4.77	1.30	-0.31 [-0.49, -0.12]
2: Replication with a new morph set	94	5.01	1.29	-0.46 [-0.60, -0.28]
3: Scale starts on right side	98	4.62	1.29	-0.26 [-0.44, -0.072]
4: Scale starts with strong intensity	92	4.63	1.17	-0.32 [-0.49, -0.13]

We then used multilevel models to predict amplification (difference between estimation and actual mean of the sequence) by their conscientiousness score. We again added a random intercept for participants.

Supplementary Table 22. The coefficients of the moderator (Big Five conscientiousness score) and participants tendency to amplify (difference between actual intensity of sequence and participants estimation) using a linear mixed model with a random intercept for participants. The coefficient shows the correlation coefficient of conscientiousness score. A one-sided t-test is used for statistical testing of the coefficients.

Study	**b [ci], (se)**	**t (df)**	**p**	**R^2**
1: Establishing effect	0.019 [-0.55, 0.59], (0.29)	0.067 (90.99)	.95	.13
2: Replication with a new morph set	-0.018 [-0.52, 0.49], (0.26)	-0.071 (92.05)	.94	.12
3: Scale starts on right side	0.90 [-0.52, 0.49], (0.22)	3.99 (95.89)	<.001***	.099

4: Scale starts with strong intensity	-0.097 [-0.49, 0.44], (0.24)	-0.097 (89.52)	.92	.067

Openness: Is being assessed by the items:

I.Open to new experiences, complex.
II.Conventional, uncreative.

Supplementary Table 23. Ns, means, and standard deviations of the Big Five personality questionnaire factor openness across Studies 1-4. As Cronbach's Alpha requires 3 or more items we show the between item correlation (*r*) instead.

Study	***n***	***M***	***SD***	***r* [ci]**
1: Establishing effect	93	4.74	1.16	-0.16 [-0.35, -0.042]
: Replication with a new morph set	94	4.90	1.11	-0.14 [-0.33, 0.058]
3: Scale starts on right side	98	4.89	1.15	-0.083 [-0.27, 0.11]
4: Scale starts with strong intensity	92	4.92	1.07	-0.17 [-0.36, 0.028]

We then used multilevel models to predict amplification (difference between estimation and actual mean of the sequence) by their openness score. We again added a random intercept for participants.

Supplementary Table 24. The coefficients of the moderator (Big Five openness score) and participants tendency to amplify (difference between actual intensity of sequence and participants estimation) using a linear mixed model with a random intercept for participants. The coefficient shows the correlation coefficient of openness score. A one-sided t-test is used for statistical testing of the coefficients.

Study	**b [ci], (se)**	**t (df)**	**p**	**R^2**
1: Establishing effect	0.35 [-0.27, 0.99], (0.32)	1.10 (91.08)	.27	.13
2: Replication with a new morph set	0.17 [-0.41, 0.77], (0.30)	0.59 (92.07)	.55	.12
3: Scale starts on right side	-0.25 [-0.78, 0.28], (0.27)	-0.92 (95.96)	.35	.099

4: Scale starts with strong intensity	-0.10 [-0.61, 0.41], (0.26)	-0.38 (89.61)	.70	.067

Positive and Negative Affect. We assessed participants positive and negative affect using a 20 item a PANAS scale (Watson et al., 1988). Participants rated how much they currently feeling certain positive and negative emotions on a scale from 1. Very slightly / not at all to 5. Extremely. Positive / Negative affect score is the average value of the 10 items corresponding to each of the emotional states.

Positive Affect: The 10 positive emotion items are:

1. *Interested*
2. *Excited*
3. *Strong*
4. *Enthusiastic*
5. *Proud*
6. *Alert*
7. *Inspired*
8. *Determined*
9. *Attentive*
10. *Active*

Supplementary Table 25. Ns, means, and standard deviations of the Positive Affect Category from the PANAS across Studies 1-4.

Study	***n***	***M***	***SD***	**α [ci]**
1: Establishing effect	93	2.50	0.82	0.87 [0.83, 0.91]
2: Replication with a new morph set	-	-	-	-
3: Scale starts on right side	98	2.52	0.74	0.86 [0.82, 0.90]
4: Scale starts with strong intensity	92	2.61	0.73	0.85 [0.81, 0.90]

We then used multilevel models to predict amplification (difference between estimation and actual mean of the sequence) by their positive affect score. We again added a random intercept for participants.

Supplementary Table 26. The coefficients of the moderator (PANAS positive affect score) and participants tendency to amplify (difference between actual intensity of sequence and participants estimation) using a linear mixed model with a random intercept for participants. The coefficient shows the correlation coefficient of positive affect score. A one-sided t-test is used for statistical testing of the coefficients.

Study	**b [ci], (se)**	**t (df)**	**p**	**R^2**
1: Establishing effect	-0.46 [-1.38, 0.45], (0.47)	-0.98 (91.01)	.32	.13
2: Replication with a new morph set	-	-	-	-
3: Scale starts on right side	0.63 [-0.18, 1.45], (0.41)	1.52 (95.91)	.13	.099
4: Scale starts with strong intensity	0.19 [-0.55, 0.94], (0.38)	0.51 (89.89)	.61	.067

Negative: The 10 negative emotion items are:

1. *Distressed*
2. *Upset*
3. *Guilty*
4. *Scared*
5. *Hostile*
6. *Irritable*
7. *Ashamed*
8. *Nervous*
9. *Jittery*
10. *Afraid*

Supplementary Table 27. Ns, means, and standard deviations of the Negative Affect Category from the PANAS across Studies 1-4.

Study	***n***	***M***	***SD***	**α [ci]**

1: Establishing effect	93	1.71	0.71	0.87 [0.84, 0.91]
2: Replication with a new morph set	-	-	-	-
3: Scale starts on right side	98	1.74	0.69	0.87 [0.83, 0.91]
4: Scale starts with strong intensity	98	1.79	0.76	0.89 [0.85, 0.92]

We then used multilevel models to predict amplification (difference between estimation and actual mean of the sequence) by their negative affect score. We again added a random intercept for participants.

Supplementary Table 28. The coefficients of the moderator (PANAS negative affect score) and participants tendency to amplify (difference between actual intensity of sequence and participants estimation) using a linear mixed model with a random intercept for participants. The coefficient shows the correlation coefficient of negative affect score. A one-sided t-test is used for statistical testing of the coefficients.

Study	**b [ci], (se)**	**t (df)**	**p**	**R^2**
1: Establishing effect	-0.38 [-1.42, 0.65], (0.53)	-0.72 (90.99)	.47	.13
2: Replication with a new morph set	-	-	-	-
3: Scale starts on right side	0.29 [-0.59, 1.18], (0.45)	0.64 (96.12)	.52	.099
4: Scale starts with strong intensity	-0.15 [-0.87, 0.56], (0.45)	-0.42 (89.86)	.67	.067

Loneliness. We assessed participants level of loneliness using the three-item loneliness scale (Hughes et al., 2004). Participants were asked to indicate how often they felt one of the three described items in their lives (from 1. Hardly ever to 3. Often). The loneliness score is the average of the three items. The three situation described by the items are:

1. *How often do you feel that you lack companionship?*
2. *How often do you feel left out?*
3. *How often do you feel isolated from others?*

Supplementary Table 29. Ns, means, and standard deviations of the Loneliness scale across Studies 1-4.

Study	***n***	***M***	***SD***	**α [ci]**
1: Establishing effect	93	1.83	0.58	0.76 [0.68, 0.84]
2: Replication with a new morph set	94	1.84	0.58	0.82 [0.76, 0.88]
3: Scale starts on right side	98	1.94	0.60	0.78 [0.70, 0.85]
4: Scale starts with strong intensity	92	1.90	0.60	0.76 [0.67, 0.84]

We then used multilevel models to predict amplification (difference between estimation and actual mean of the sequence) by their loneliness score. We again added a random intercept for participants.

Supplementary Table 30. The coefficients of the moderator (loneliness score) and participants tendency to amplify (difference between actual intensity of sequence and participants estimation) using a linear mixed model with a random intercept for participants. The coefficient shows the correlation coefficient of loneliness score. A one-sided t-test is used for statistical testing of the coefficients.

Study	**b [ci], (se)**	**t (df)**	**p**	**R^2**
1: Establishing effect	-0.18 [-1.45, 1.08], (0.64)	-0.28 (90.95)	.77	.13
2: Replication with a new morph set	0.14 [-0.98, 1.27], (0.57)	0.24 (91.98)	.80	.12
3: Scale starts on right side	0.11 [-0.88, 1.11], (0.51)	0.22 (94.96)	.82	.099
4: Scale starts with strong intensity	0.46 [-0.45, 1.37], (0.46)	0.90 (89.74)	.32	.067

Comparison: Peak-End Rule vs. Amplification Effect

The goal of this section was to compare the prediction made by the peak-end rule and participants' estimation of the mean. According to the peak-end rule the average of the most intense face of a sequence and the intensity of last face displayed is the best estimate for people's evaluation of the sequence. To compare the prediction made by the peak-end to participants' actual estimation, we compared participants' estimation of the sequence to both the actual mean intensity and the prediction made by the peak-end rule. For this comparison we used a mixed model analysis of repeated measures, comparing the participants' estimated mean emotion with the actual mean emotion expressed in each set and the score as predicted by the peak-end. We added a by-participant and by-face-identity random intercepts (Supplementary Tables 31 – 34, Supplementary Figures 5 – 8). Results show that the peak-end was significantly larger than the estimation by the participants.

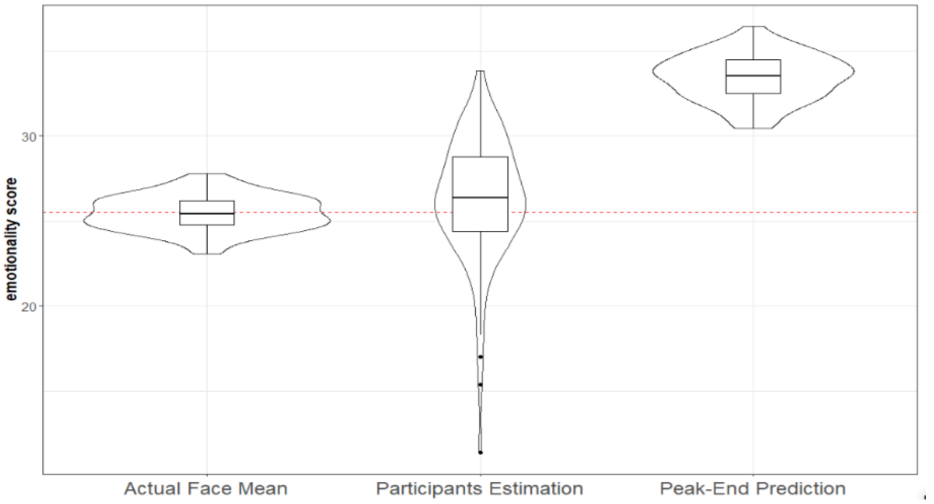
Study 1 Actual Face Mean compared to Estimation and Peak-End Prediction

Supplementary Table 31. Linear mixed model comparing the value of three levels (baselevel: estimation value, comparison levels: peak-end prediction, actual face mean). We used a participant-id as a random intercept for the model. A one-sided t-test is used for statistical testing of the coefficients.

Fixed Effects					
	Estimate	*SE*	95% CI	*t*	*p*
Estimation	26.24	0.19	25.86 – 26.62	135.62	.000
Face-Mean	-0.75	0.18	-1.11– -0.39	-4.12	.000
Peak-End	7.20	0.18	6.84– 7.56	39.31	.000
Random Effects					
			Variance	*SD*	
Participant (Intercept)			1.92	1.38	
Residual			82.21	9.06	
Model Fit					
R^2			Marginal	Conditional	
			0.13	0.15	

Model equation: Value ~ ValueType + (1|Participant_id)

Notes. Model fit was calculated using the R package MuMIn[7] based on the paper by Nakagawa and colleagues[8].



Supplementary Figure 5. Violin plots showing the comparison of the actual face mean of a sequence to participants estimation and the prediction based on the peak-end rule. The box plots display the median, first, and third quartiles. The total number of participants is n = 93. The whiskers extend to the most extreme value less than 1.5 times the interquartile range beyond the quartile. Outliers, which are data points outside the 1.5 interquartile range, are plotted as individual dots. Red dotted line represents the middle of the emotional intensity of the stimuli. The result shows that the prediction made by the peak-end rule is sig larger than the actual estimation of the participants.

Study 2 Actual Face Mean compared to Estimation and Peak-End Prediction

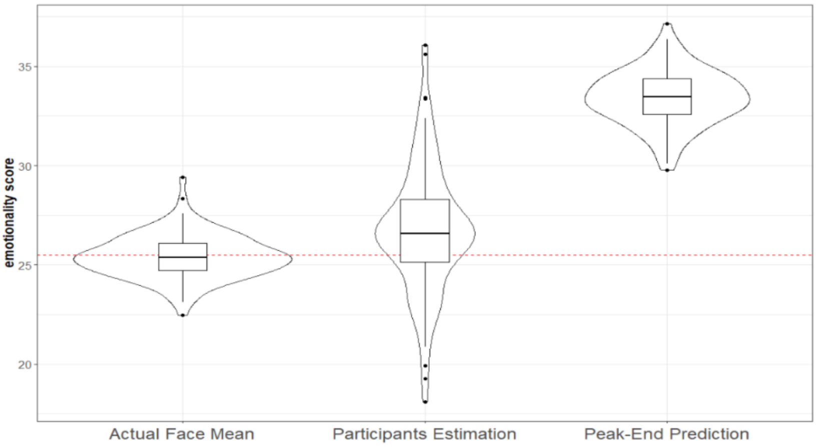
Supplementary Table 32. Linear mixed model comparing the value of three levels (baselevel: estimation value, comparison levels: peak-end prediction, actual face mean), using a random intercept of participants' id. A one-sided t-test is used for statistical testing of the coefficients.

	Fixed Effects				
	Estimate	*SE*	95% CI	*t*	*p*
Estimation	26.79	0.19	26.42 – 27.15	144.25	.000
Face-Mean	-1.37	0.18	-1.72 – -1.01	-5.77	.000
Peak-End	6.62	0.18	6.27 – 6.98	36.49	.000
	Random Effects				
			Variance	*SD*	
Participant (Intercept)			1.70	1.30	
Residual			74.02	8.60	

Model Fit		
R^2	Marginal	Conditional
	0.14	0.16

Model equation: Value ~ ValueType + (1|Participant_id)

Notes. Model fit was calculated using the R package MuMIn (Barton, 2018) based on the paper of Nakagawa et al. (2017).



Supplementary Figure 6. Violin plots showing the comparison of the actual face mean of a sequence to participants estimation and the prediction based on the peak-end rule. The box plots display the median, first, and third quartiles. The total number of participants is n = 94. The whiskers extend to the most extreme value less than 1.5 times the interquartile range beyond the quartile. Outliers, which are data points outside the 1.5 interquartile range, are plotted as individual dots. Red dotted line represents the middle of the emotional intensity of the stimuli. The result shows that the peak-end prediction is significantly larger than the actual estimation of the participants.

Study 3 Actual Face Mean compared to Estimation and Peak-End Prediction

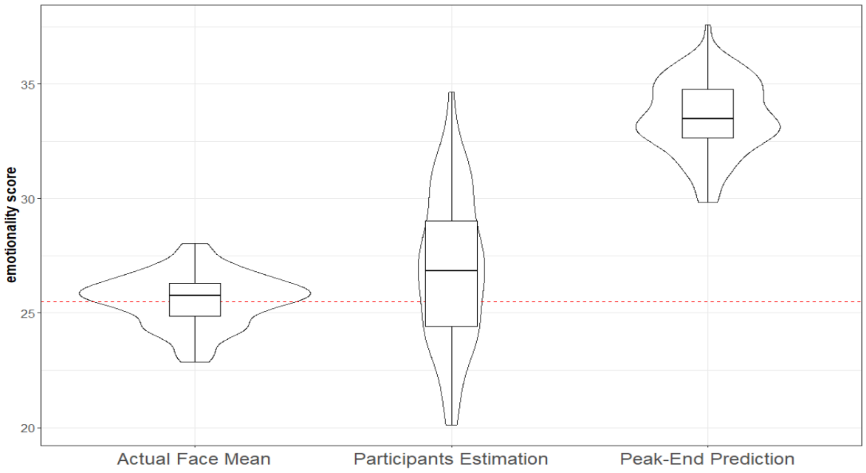
Supplementary Table 33. Linear mixed model comparing the value of three levels (baselevel: estimation value, comparison levels: peak-end prediction, actual face mean). A one-sided t-test is used for statistical testing of the coefficients.

Fixed Effects					
	Estimate	*SE*	95% CI	*t*	*p*
Estimation	26.91	0.19	26.54 – 27.28	143.88	.000
Face-Mean	-1.31	0.18	-1.67 – -0.96	-7.33	.000

Peak-End	6.62	0.18	6.26 – 6.97	36.79	.000
Random Effects					
			Variance	*SD*	
Participant (Intercept)			1.84	1.35	
Residual			81.87	9.05	
Model Fit					
R^2			Marginal	Conditional	
			0.13	0.15	

Model equation: Value ~ ValueType + (1|Participant_id)

Notes. Model fit was calculated using the R package MuMIn (Barton, 2018) based on the paper of Nakagawa et al. (2017).



Supplementary Figure 7. Violin plots showing the comparison of the actual face mean of a sequence to participants estimation and the prediction based on the peak-end rule. The box plots display the median, first, and third quartiles. The total number of participants is n = 98. The whiskers extend to the most extreme value less than 1.5 times the interquartile range beyond the quartile. Outliers, which are data points outside the 1.5 interquartile range, are plotted as individual dots. Red dotted line represents the middle of the emotional intensity of the stimuli. The result shows that the prediction made by the peak-end rule is significantly larger than the actual estimation of the participants.

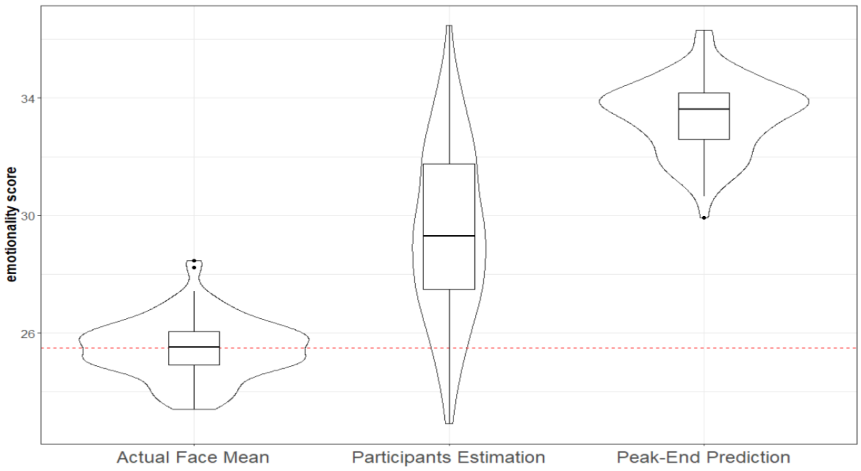
Study 4 Actual Face Mean compared to Estimation and Peak-End Prediction

Supplementary Table 34. Linear mixed model comparing the value of three levels (baselevel: estimation value, comparison levels: peak-end prediction, actual face mean). A one-sided t-test is used for statistical testing of the coefficients.

Fixed Effects					
	Estimate	*SE*	95% CI	*t*	*p*
Estimation	29.43	0.17	29.08 – 29.77	167.16	.000
Face-Mean	-3.95	0.19	-4.32 – -3.58	-21.16	.000
Peak-End	3.94	0.19	3.57 – 4.30	21.09	.000
Random Effects					
			Variance	*SD*	
Participant (Intercept)			1.24	1.11	
Residual			82.87	9.10	
Model Fit					
R^2			Marginal		Conditional
			0.11		0.12

Model equation: Value ~ ValueType + (1|Participant_id)

Notes. Model fit was calculated using the R package MuMIn (Barton, 2018) based on the paper of Nakagawa et al. (2017).

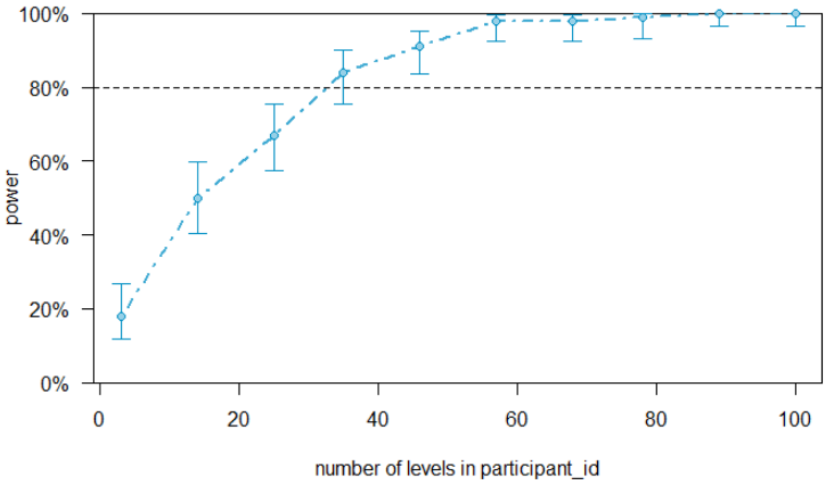


Supplementary Figure 7. Violin plots showing the comparison of the actual face mean of a sequence to participants estimation and the prediction based on the peak-end rule. The box plots display the median, first, and third quartiles. The total number of participants is n = 92. The whiskers extend to the most extreme value less than 1.5 times the interquartile range beyond the quartile. Outliers, which are data points outside the 1.5 interquartile range, are plotted as individual dots. Red dotted line represents the middle of the emotional intensity of the stimuli. The result shows that the prediction made by the peak-end rule is significantly larger than the actual estimation of the participants.

3. Study 5: Testing the Role of Memory for the Amplification Effect

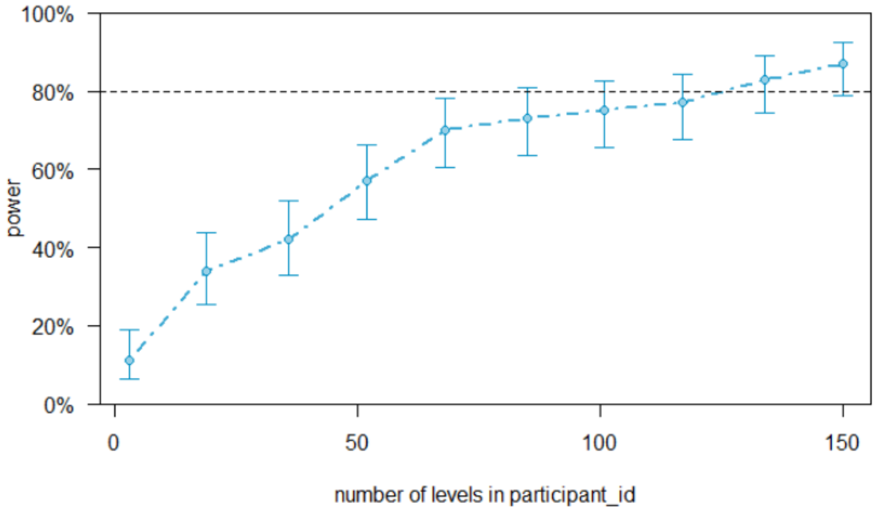
Power Analysis to Determine Sample Sizes

As Study 5 was meant to consist of trials using 8 faces per sequence, we conducted our power analysis for Study 5 on a subset of Study1, looking only at the trials of 8 faces per sequence. We used this estimation difference to predict the power using the simr package [9]. The power analysis showed that running a study with 100 participants would result in a power of nearly 100% to find the amplification effect in 20 trials (Supplementary Figure 9).



Supplementary Figure 9. Power-analysis of the amplification block for study 5. Data are presented as the mean power estimation and the error bars represent bootstrapped 95% CIs based on 2,000 simulations. This estimation was based on results from Study 1, specifically looking only at sequences of 8 faces, given that in Study 5 these would be the sequences that participants would see. This resulted in a sample of n = 92 participants who completed 856 trials in total which showed 8 faces for the power calculation. The power analysis showed that having 100 participants doing 20 trials results in a power of nearly 100%.

After running a first version of Study 5 with only 20 trials containing a sequence of 8 faces, we found that the amplification effect was much smaller when the sequence length did not vary. The actual observed power was 60% as estimated by the simr package. As a comparison, the difference between the estimated sequence and the actual sequence was down from a difference of 1.07 to .50. To create a high powered study, we then used this study's effect size to update our power analysis and pre-registration. Our simulation suggested that increasing the trial number to 30 and participants number to 150 would lead to power of 87% (Supplementary Figure 10).



Supplementary Figure 10. Power-analysis of the amplification block for study 5. Data are presented as the mean power estimation and the error bars represent bootstrapped 95% CIs based on 2,000 simulations. This estimation was based the estimation difference of the previous attempt of this study. A sample of n = 83 participants who 20 trials each was used in the power calculation. The power analysis showed that having 150 participants doing 30 trials results in a power of approximately 86%.

Detailed Results

We started our analysis by looking at the first block in which participants were asked to evaluate the average intensity of the sequence. To measure general tendency for amplification, we conducted a mixed model analysis in which we compared participants' estimation of the sequences to the actual average, using by-individual and face-identity random intercepts. Similar to Studies 1a-d and as predicted in Hypothesis 1, results suggested that participants evaluated the sequence as more emotional than it actually was (b = 0.59, t(8,846) = 4.02, p < .001, R^2= .10, 95% Confidence Intervals = [0.30, 0.89]), although the effect was smaller than those found in Studies 1a-d. This may be due to the fact that sequence length didn't change which made it easier for participants to improve on the task, which is also evident by the lower SE compared to Studies 1a-d (.14 in this study compared to .16 -.17 in Studies 1a-d) despite having less trials than those studies. We then compared the amplification of negative and positive sequences. This was done by looking at the difference score between participant's estimation of the average sequence as the dependent variable and using valence as the independent variable. Similar to previous models, the model used participant-id and face-identity as random intercepts. As predicted in Hypothesis 2, results suggested that amplification of negative sequences was significantly stronger than positive sequences (b = 1.12, t(8,849) = 4.72, p < .001, R^2= .26, 95% Confidence Intervals = [0.65, 1.59]).

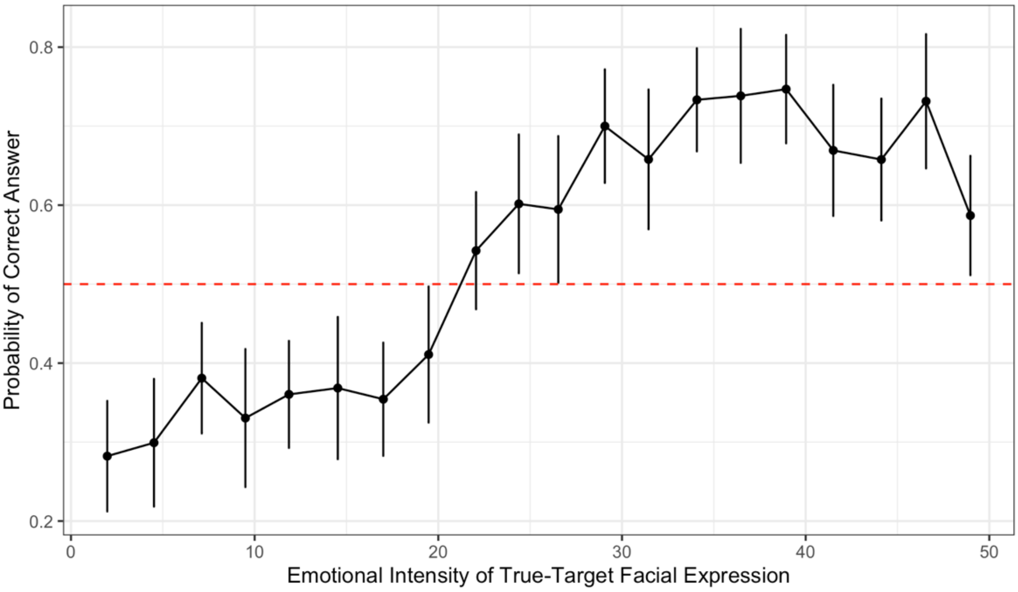
After confirming the general effect of amplification, we examined the weight given to each of the expressions in the sequence. This was done in a similar way to Hubert and colleagues [10], by using each expression number as a separate coefficient in a mixed model regression predicting participants' estimation. We removed the intercept of the model to get a comparison of each coefficient to zero as well as added a by-participant and face-identity as random intercepts, similar to all other analyses. See the left size of Supplementary Table 35 (Expression Number Predicting Estimation) for the size of each coefficient. As expected, and aligned with previous findings, recent expressions in the sequence were stronger predictors of participants' estimation of the sequence average, which would suggest that they were more salient in participants' memory.

Supplementary Table 34. Results for the two analyses that examined the order of the expressions in the sequence predicting either participants' estimation of the average sequence emotionality (left) or whether participants were able to recall the expression in the memory test (right).

Expression Location	**Expression Location Predicting Estimation**		**Expression Location Predicting Correct Memory**	
	***b* [ci],**	***t***	***b* [ci]**	***z***
1	.04 [.02, .06]	5.80***	-.04 [-.23, .15]	-.44
2	.04 [.02, .06]	5.59***	-.01 [-.20, .17]	-.18
3	.05 [.03, .06]	6.61***	-.11 [-.08, .31]	-1.16
4	.05 [.04, .07]	7.46***	-.09 [-.30, .11]	-.91
5	.07 [.05, .08]	9.45***	.12 [-.07, .32]	1.18
6	.07 [.05, .08]	8.87***	-.03 [-.22, .16]	-.32
7	.08 [.06, .09]	10.48***	.31 [.12, .50]	3.20**
8	.011 [.09, .012]	14.16***	.53 [.33, .74]	5.16***

Having established both amplification and recency effects, we then turned to the memory block, examining whether the emotional intensity of the target facial expression predicted the probability of remembering the expression (Hypothesis 3). To evaluate this question, we conducted a mixed generalized linear model in which we used the emotional intensity of the true target expression as predictor and whether participants chose this expression correctly or not as the dependent variable. We added a covariate to the model of the distance between the false and the true target, as such distance is likely to affect participants ability to remember. We also added a by-individual random intercept and a random intercept of the face identity. As hypothesized, results suggested that the intensity of the facial expression predicted the probability of memory (b = 0.04, z = 15.23, p < .001,

R^2 = .10, 95% Confidence Intervals = [0.037, 0.048], Supplementary Figure 11). To make sure that this effect was not driven solely by participants merely choosing the more emotional expression in each trial, we reduced the dataset to the cases in which the true-target expression was lower in intensity than the false target expression and redid the analysis. The effect remained significant within this subset despite cutting the sample size by half (b = 0.01, z = 2.16, p = .03, R^2 =.030, 95% Confidence Intervals = [0.001, 0.029]). We did not find a significant main effect for valence on participants' memory as well as an interaction between the intensity of the true-target stimuli and its valence on the probability of memory.



Supplementary Figure 11. Emotional intensity of the true-target facial expression predicting the probability for correct answer. Data are presented as mean values +/- SE. Sample size is n = 150. We aggregated the data into 20 slots in order to provide an estimate of the standard error of the predictions for the sake of visualization. Red dotted line represents chance accuracy.

In addition to testing whether emotional salience predicted memory, we also wanted to replicate the recency findings from the first block by examining whether true-target expressions were more likely to be identified if they appeared later in the sequence. To evaluate this question, we created a binary value for each of the 8 facial expressions that participants saw in each trial, zero indicating that a specific expression was not the true target expression and 1 indicating that it was. For most trials, there was only one value in the sequence that was similar to the true-target expression, but in 13.43% of cases the true target expression corresponded to two or more expressions in the sequence. We then conducted a mixed model analysis in which we used each of these binary values for each facial expression as the independent values, and whether participants were correct in the memory test as the dependent variable. The model included by-individual and by-face-identity random variables and we removed the intercept of the model to evaluate the importance of each expression location, similar to the model that attempted to evaluate the importance of location on amplification. As seen in the right side of Supplementary Table 35, the order of the true-target expressions in the sequence predicted whether the memory was correct only at the 7th or 8th order. This finding provides some support to the notion of recent items being more salient in memory. In addition, these findings can be seen as providing additional evidence to the memory task was quite difficult.

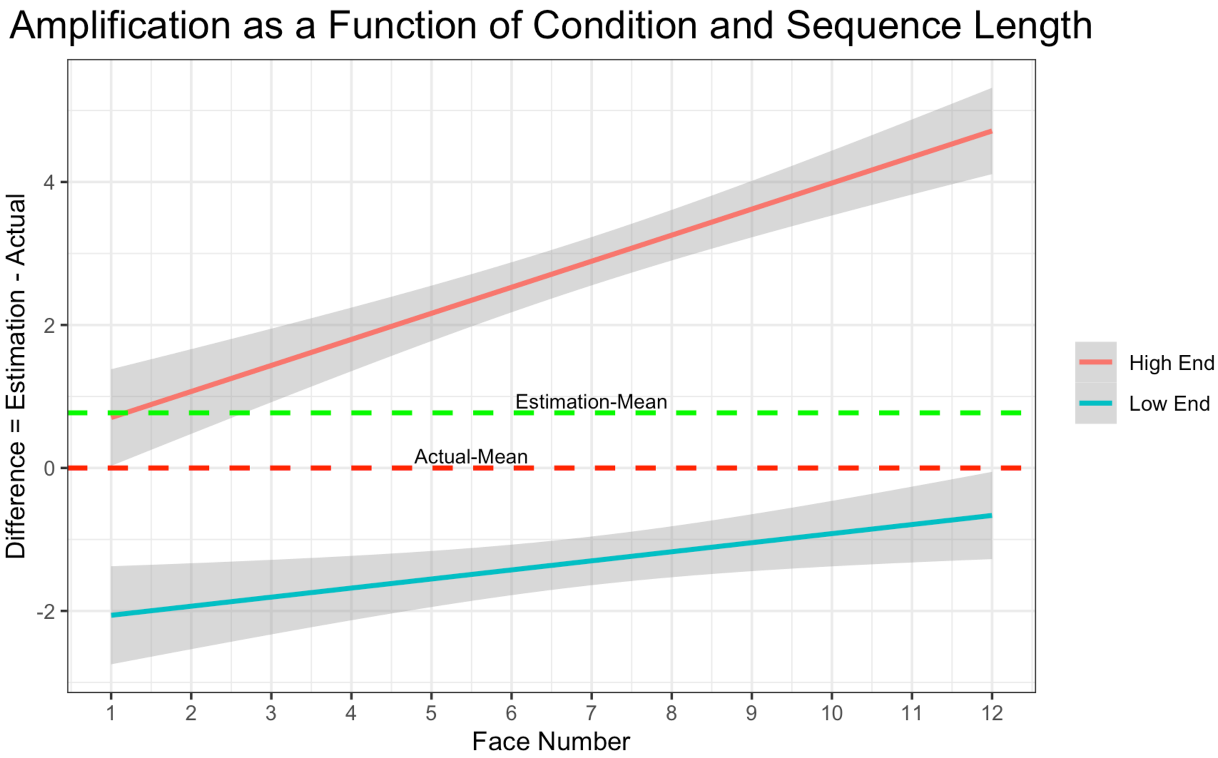
One exploratory hypothesis that was proposed in our pre-registration was that individuals' tendency to correctly remember stronger stimuli could potentially predict tendencies for amplification. To examine this possibility, we created a coefficient for each participant of facial emotional intensity predicting memory: specifically, we ran a generalized logistic regression for each participant looking at emotional intensity of the true target expression as a predictor the probability of memory (emotional intensity predicting memory). We then took the regression coefficient for each participant, such that a positive coefficient indicated that, for that participant, increased intensity predicted better memory. The intensity-memory coefficients were then used to predict tendency for amplification in the sequence evaluation block. This was done by running a mixed model analysis using the participant-level memory coefficient as a predictor and the difference between the estimated sequence intensity and the actual intensity in each trial. Similar to previous analyses, we added by-individual and by-face-identity random intercepts. Results suggested that the association between the memory-intensity coefficient and participants' tendency for amplification was non-significant (b = 5.94, t (147) = .83, p = .40, R^2 = .24, 95% Confidence Intervals = [-8.12, 20.02]). This means that individual level tendencies for memory did not predict trial level amplification.

4. Study 6: Testing the Impact of Recency and Emotionality on Participants' Evaluation

Detailed Results

We first tested the three amplification hypotheses as we had done in Studies 1-4, using identical analyses. Results suggested that participants generally estimated the emotions in the sequence as more intense than they actually were (b = 0.77, t (9,814) = 5.89, p < .001, R^2= .06, 95% Confidence Intervals = [0.51, 1.02]). Our tests of Hypotheses 2 and 3 indicated that the length of the sequence also led to increased amplification (b = 0.25, t (4,90) = 7.34, p < .001, R^2= .14, 95% Confidence Intervals = [0.18, 0.32]), and that negative sequences led to a significantly stronger amplification than positive sequences (b = 0.69, t (4,895) = 2.92, p < .001, R^2= .13, 95% Confidence Intervals = [0.19, 1.13]).

Next, we evaluated the tendency for amplification in the high-intensify-end and the low-intensity-end trials. This was done by looking at the difference between participants' estimation and the actual sequence average as the dependent variable, and the order of high and low intensity expressions as the independent variable, including a by-participant and by-face-identity random variables (see Supplementary Figure 12). Looking first at the intercept of the model, which is the high intensity end condition, results suggested that in trials in which the high intensity expressions were presented at the end, participants estimations were significantly higher than zero (b = 2.49, t (23.64) = 4.73, p < .001, R^2= .19, 95% Confidence Intervals = [1.45, 3.52]). Results also suggested that amplification was significantly higher in the high intensity end trials compared to the low intensity end trials (b = -4.19, t (4,895) = -18.29, p < .001, R^2= .19, 95% Confidence Intervals = [-4.65, -3.74]). To further understand the degree of amplification/de-amplification in the low intensity end condition we releveled the conditions to examine the intercept of the model which represent the difference between the low-intensity-end trials to zero. Results pointed to a significant de-amplification in the low intensity end condition (b = -1.31, t (4,895) = -2.90, p = .01, R^2= .19, 95% Confidence Intervals = [-2.24, -0.38]). There was no significant interaction between valence and the manipulation of order.



Supplementary Figure 12. A summary of Study 6 results. The y-axis represents the difference between participants' estimated mean and the actual mean. Data are presented as mean value +/- standard errors. Sample size is n = 100. Corresponding to this, the green dotted line represents the average of participants' estimated mean while the red dotted line represents the estimated mean. The x-axis represents the number of facial expressions that participants saw in each trial. Finally, the red or blue lines represent whether the high-intensity expressions or low-intensity expressions were at the end of the sequence.

5. Study 7: Manipulating Salience

The goal of Study 7 was to examine the effect of memory on participants' tendency for amplification by manipulating the salience of either high or low intensity emotions. Manipulation of salience was done by adding a red square around either the high intensity or low intensity expressions in a sequence. Our pre-registered hypotheses (https://osf.io/yxqem/) were that we would see amplification (H1), that amplification would increase with sequence length (H2) and that amplification would be stronger for negative emotions (H3). Finally, we hypothesized that amplification would be higher for the task in which high intensity expressions were emphasized by our salience manipulation compared with when low-intensity expressions were emphasized. We had no clear prediction regarding whether amplification in the control condition (no salience) would be more similar to results from the high or low intensity salience conditions (H4).

Methods

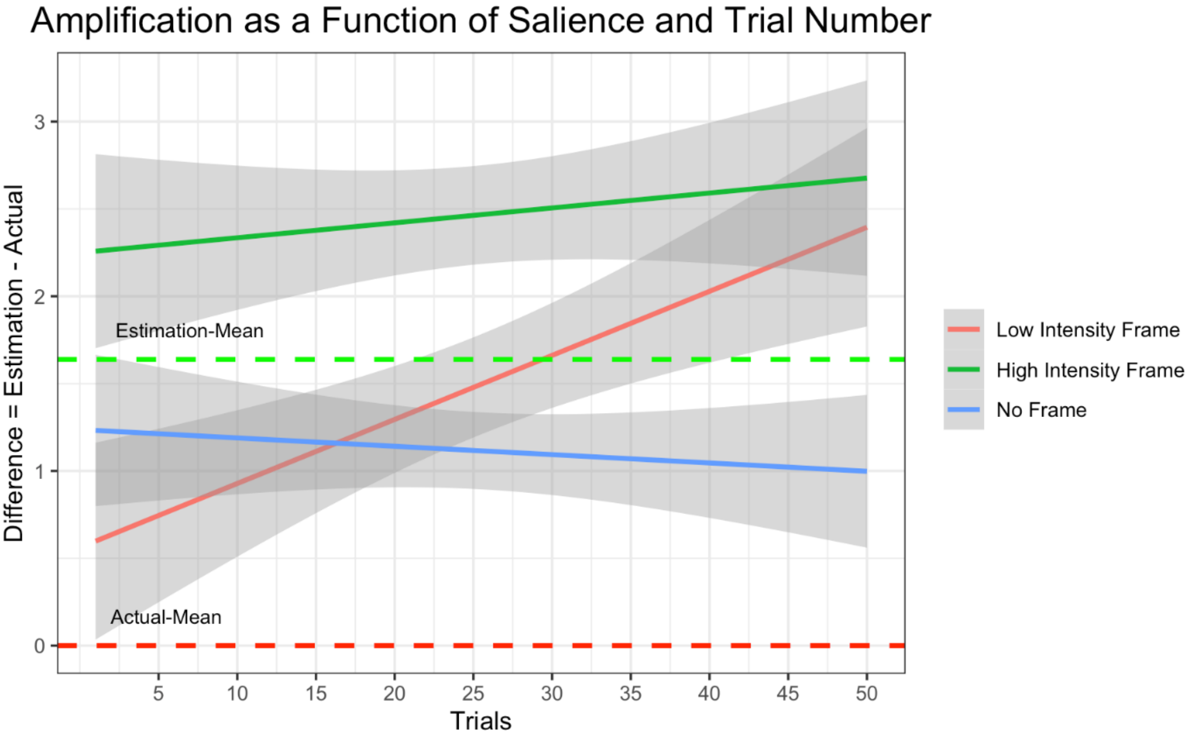
Participants. Unlike our order manipulation in Study 6, which was likely undetected by participants, we worried that our salience manipulation would be obvious to participants if we employed a within-participants design. We therefore decided to use a between-subjects design and recruited 100 participants per each condition for a total N = 300. Participants were recruited from prolific in exchange for $2.30. Informed consent was obtained by participants. Our initial sample was N= 300. Congruent with our pre-registered criteria, we removed 5 participants for providing average ratings of below 10 or above 40. Our final sample was therefore N = 295 (men: 119, women:175, other: 1, Age: *M* = 25.14, *SD* = 8.06).

Procedure. Participants were randomly assigned to one of three conditions (between-subject design): high salience, low salience, and control. In the high salience condition, participants completed a task that was similar to that of Studies 1a-d with one difference: every time a facial expression with emotional intensity of more than 29 was presented, this expression was presented with a red frame around it. We chose a threshold of 30 and not 25 to slightly reduce the frequency of red frames and increase the potential salience. In the low salience condition, red frames appeared around expressions with an emotional intensity less than 21. The control condition was identical to Study 1a. Following the task, participants completed a survey similar to previous studies (see SI for full analysis).

Results

We first tested the three amplification hypotheses using analyses identical to those used in Studies 1a-d. Results suggested that participants generally estimated the emotions in the sequence as more intense than they actually were (b = 1.63, t (429,023) = 16.49, p < .001, R^2= .07, 95% Confidence Intervals = [-1.44, 1.83]). The length of the sequence also led to increased amplification (b = 0.36, t (14,406) = 18.09, p < .001, R^2= .22, 95% Confidence Intervals = [0.32, 0.39]). Finally, negative sequences led to a significantly stronger amplification than positive sequences (*b* = 0.52, *t* (14,409) = 3.69, *p* < .001, R^2= .20, 95% Confidence Intervals = [0.25, 0.79]).

Next, we compared the magnitude of amplification between the three conditions, low intensity salience, high intensity silence, and control. We used our control condition as a baseline condition. Results suggested that the high salience condition led to significantly stronger amplification compared to the control condition (*b* = 1.39, *t* (291) = 2.44, *p* = .01, R^2= .19, 95% Confidence Intervals = [0.28, 2.51]). There was no significant difference between the control condition and the low salience condition (*b* = 0.40, *t* (291) = 0.69, *p* = .48, R^2= .19, 95% Confidence Intervals = [-0.73, 1.54]). One potential reason for the lack of difference between the low salience and the control condition may be that the salience of the emotional expressions overrode the manipulation, and so once participants got used to the red frames, they returned to focus on the high intensity expressions. To evaluate this possibility in an exploratory analysis, we tested for an interaction between the condition and the trial number. Looking first at the interaction between control and high salience, results suggested that the interaction was non-significant, such that there was no difference in the degree of amplification between the control and the high silence conditions (*b* = 0.01, *t* (14,365) = 1.44, *p* = .14, R^2= .19, 95% Confidence Intervals = [-0.005, 0.038]). However, there was a significant interaction between trial and condition when comparing the control condition and the low salience condition (*b* = 0.04, *t* (14,366) = 3.81, *p* < .001, R^2= .19, 95% Confidence Intervals = [0.02, 0.06]). Further analysis of simple effects of the amplification over trial number for each of these two conditions by centring the model on different conditions suggested that there was no significant change in amplification in the control condition (*b* = -0.006, *t* (14,366) = -0.86, *p* = .38, R^2= .19, 95% Confidence Intervals = [-0.02, 0.01]). However, for the low salience condition, amplification significantly increased with trial number (*b* = 0.03, *t* (14,365) = 4.23, *p* < .001, R^2= .19, 95% Confidence Intervals = [0.02, 0.05], see Supplementary Figure 13). While these results do not provide direct evidence as to whether emotional salience led participants in the low salience condition to focus more on emotional expressions as the task progressed, it is congruent with the notion that there was a change in the way that participants in the low salience condition were completing the task as it unfolded, while no such change was seen in the other conditions.



Supplementary Figure 13. A summary of Study 7 results. The y-axis represents the difference between participants' estimated mean and the actual mean. Corresponding to this, the green dotted line represents the average of participants' estimated mean while the red dotted line represents the estimated mean. Data are presented as mean value +/- standard errors. The total sample size is n = 300. The x-axis represents the trials number. Finally, the red, green and blue lines represent whether the salience manipulation of the red frames was done on the high intensity frame, low intensity trials or control condition.

Taken together, our manipulation of salience was successful in differentiating the high salience and the control condition, such that salience increased amplification. However, no difference was found in degree of amplification between the control and the low salience condition. This lack of differentiation could have been caused by the salience manipulation being overwhelmed by the natural salience of the high intensity expressions. Exploratory analyses looking at amplification as a function of trial number in the low salience condition indicated that amplification increased with trial number, pointing to the possibility that as participants in the low salience condition habituated to the salience of red frames, the effect of high intensity emotions became more salient.

6. Study 8: Examining Amplification Over and Above Nonlinearity in Emotion Perception

Findings from this study series support the idea that when perceivers estimate the mean emotion of a sequentially presented series of emotion expressions, they systematically over-estimate the mean. Thus far, our analysis of underlying mechanisms has focused on differential memory for more emotional (intense, salient) stimuli. However, another explanation of the apparent bias we observe is that it arises from a nonlinearity in the perception of the emotional facial expressions that we presented. More precisely, it is possible that there is a perceptual asymmetry in how our lower-valence expressions are perceived relative to our higher valence expressions, such that a middle-valence expressions are seen as more similar to higher-valence expression than to the lower-valence expressions. Participants would then be more likely to confuse the middle-valence expressions with the higher-valence ones than the lower valence ones, leading to an amplification in their estimation of the average. This explanation would entail that it is the perceptual characteristics of the stimulus space, rather than changes in memory that drive the observed amplification effects. Given this concern, we used a computational modeling approach to separately quantify the psychophysical similarity between expressions, and used this similarity data to estimate what biases in memory for ensembles would be expected based on similarity alone.

To achieve this goal, we first empirically tested how people perceived distances between emotional intensities at different points of our emotional scale. We then built on an existing computational model (Robinson & Brady, 2021) that was designed to simulate ensemble memory with specific attention to non-linearity in similarity, by comparing three hypothetical models of ensemble coding: A baseline model, that only incorporated nonlinearity in similarity, a recency model that was based on the baseline model but also assumed stronger weight in memory for more recent items, and an amplification model that was based on the recency model but added an assumption of increased weight to more emotional expressions. We used the results of Study 6 to compare these three models' fit.

Method

The current project involved two steps. In the first step, we ran a similarity task to examine non-linearity in participants' estimation of similarity at different locations of the scale. In the second step we implemented the findings from our similarity study and compared the fit of three hypothetical models using the data from Study 6.

Participants. We recruited participants from Prolific in exchange for $2.70. Informed consent was obtained by participants. We aimed for a similar number of participants as in our other studies. No participants were excluded from the study. Our final sample was N= 100 (men: 37, women:62, other: 1, Age: *M* = 35.99, *SD* = 12.69).

Procedure. Recent work by Schurgin, Wixted and Brady (2020) took a computational modeling approach to delineate the effects of the psychophysical similarity of stimuli on well-known memory phenomena in visual memory. These authors found that once the psychophysical similarity of stimuli was taken into account, many purported memory phenomena were in fact reducible to the perceptual confusability of stimuli. To evaluate similarity in emotional perception we modified the similarity task that was used by Schurgin and colleagues. In each trial, participants saw two expressions on the screen and were asked to evaluate to what degree these two expressions were similar to each other on a 1-7 scale, 1 – not similar at all, 7- very similar. Participants had as much time as they needed to make their selection The similarity between two expressions was measured using a seven-point Likert scale, where Smin= 1 and Smax = 7. To generate the psychophysical similarity function, we simply normalized these data to range from 0 to 1, giving a psychophysical similarity metric, such that f(x) = ((Sx− Smin)/(Smax − Smin)). In order to cover the whole 1-50 scale, one of every five expressions was selected and compared to all other expressions in increments of 5. For example, an expression of emotional intensity 1 was compared to: 1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50. Completing all comparisons within a certain scale required conducting 66 comparisons. In each study participants completed 264 (66 X 4) comparisons, which meant that each participant completed all possible comparisons in four out of the eight expression-emotion continua: 4 identities X 2 valences (neutral-to-happy and neutral-to-angry). The 4 identities that were chosen randomly for each participant.

Results

Similarity Analysis. Our first analysis involved evaluating the linearity in the relationship between the actual distance between the expressions and the estimated distance. We conducted a mixed model analysis looking at actual item similarity predicting perceived similarity. To evaluate non linearity we introduced a second and third-order polynomial in addition to the linear slope of the model using the poly function in r [12]. Results suggested that both the linear term (*b* = -0.24, *t* (26,297) = -85.72, *p* < .001, 95% Confidence Intervals = [0.28, 2.51]), the second-order polynomial (*b* = -0.006, *t* (26,297) = -2.77, *p* = .01, 95% Confidence Intervals = [0.28, 2.51]), and the third-order polynomial (*b* = 0.02, *t* (26,297) = 15.02, *p* < .001, 95% Confidence Intervals = [0.28, 2.51]) were significantly associated with the data. Model comparisons of the current model (AIC =-1279.92) with both a model which had only a linear and a second-degree polynomial slope (AIC = -1068.48) and only a linear model slope (AIC =-897.63), suggested that the model with the third-degree polynomial slope was the strongest.

Comparing Ensemble Coding Models. Having established non-linearity in similarity perceptions, we then took a computational modeling approach to validate that indeed that the amplification found in our studies did not stem from nonlinearity and perception of similarity. We adapted a recently developed model for ensemble memory (Robinson & Brady, 2021), which is the first computational model to make high-precision predictions of performance in continuous report memory ensemble tasks. In this work, we treat this memory for ensembles as a measurement model; that is, as explained, we use it to formally separate effects of psychophysical similarity from amplification memory biases.

This model of memory for ensembles postulates that each stimulus evokes a distributed pattern of activation over feature values, and ensembles are computed by pooling over these patterns of activation at a relatively early perceptual stage of processing. Critically, within this modeling framework, the pattern of activation evoked by each stimulus depends on the psychophysical similarity of features to items held in memory, such that feature values that are more like items held in memory receive a higher boost in activation. This model directly links psychophysical similarity to memory processes by postulating that the patterns of activation elicited by each stimulus determines how familiar that feature, and similar features will feel. For instance, if a task requires remembering a certain emotional intensity, the specific intensity will evoke a very strong familiarity signal, but so will similar emotional expressions. Finally, in line with mainstream signal detection models of memory [13], the model posits that ensemble memory representations are corrupted by noise and that the signal-to-noise ratio depend on factors that determine the top-down upweighting of features of individual items (e.g., manipulation of memory load, delay, or presentation format). Formally, the most straightforward version of this model for ensembles is given by the following equation:

$$R_{ENS} = argmax\left(\left(\sum_{i=1}^{N} f(x)_i \, d'\right) + \sigma_{Noise}\right) \qquad (1)$$

where R_{ENS} is the reported feature on the ensemble task (i.e., which expression is chosen), N is the total number of items in the ensemble memory array, $f(x)$ is the psychophysical similarity function of item i (i.e., it captures how similar each of the 50 expressions are to item i; we describe the measurement of this below). d' is a free parameter that determines the level of activation of each feature value for each item. Note that this version of the model postulates that on average, each item in the sequence generates the same familiarity signal, meaning that d' is the same value for each item in the sequence (that is, the model only has one free parameter d'). σ_{Noise} is a fixed amount of noise[1,] and $argmax$ denotes the decision rule that memory reports are based on the feature that generates the maximum familiarity signal. More precisely, the argmax argument is taken over a vector of random variables ($X_{1,}X_{2,}X_{3,}\ldots X_{50}$), where each random variable is one of the fifty possible expressions on the self-report scale, each of which is distributed according to the model equation given in the parentheses. We refer to the above model as the Baseline model because it assumes that 1) the familiarity of the ensemble is solely determined by its psychophysical similarity, and that on, on average, 2) there is equal weighting of each item in memory – that is, there are no sequential or amplification effects on memory (i.e., no recency or exaggeration of the impact of negative expressions).

The second variant of the ensemble model we use is the Recency model (Robinson & Brady, 2021), which postulates that memory performance in the sequential paradigm is determined by psychophysical similarity as well as higher weighting of more recent items in memory (recency

[1] Noise is set to one standard deviation of a Gaussian distribution, consistent with a signal detection model.

effects). In line with extant recency models of memory, the recency weights are quantified with a normalized exponential function (without base e) defined over the serial position of each stimulus in the sequence [14]. The recency model is given by the following equation:

$$R_{ENS} = argmax\left(\left(\sum_{i=1}^{N} f(x)_i\, d'\, w_i^{Recency}\right) + \sigma_{Noise}\right) \quad (2)$$

$$w_i^{Recency} = \frac{r^i}{\sum_{i=1}^{N} r^i} \quad (3)$$

where $w_i^{Recency}$ is the recency weight of the ith item in the sequence, and r is a free parameter that the rate of prioritization as a function of the serial position of a stimulus (Tong et al., 2019). This version of the model, therefore, has two free parameters, d' and r. The critical point to note is that Equations 1 and 2 are identical except that Equation 2 can also capture higher weighting of more recent items. Thus, a comparison of these models provides insight into whether there is evidence for higher prioritization of more recent items in the sequence once psychophysical similarity is taken into account. Given prior evidence for recency effects in ensemble tasks, as well as the studies reported above, we expect the recency model to outperform the baseline model.

The final model we refer to as the Amplification Recency model. This model of ensembles postulates that in addition to effects of psychophysical similarity and recency on memory, there is also amplification of emotional expressions. As noted, we use this model as a measurement model[2], to formally separate possible effects of amplification from psychophysical similarity and recency. Accordingly, in line with our behavioral results, we make the simplifying assumption that recency and amplification combine independently to bias memory, and that amplification increases exponentially as a function of an expression's emotional extremeness. The model equation is shown below.

$$R_{ENS} = argmax\left(\left(\sum_{i=1}^{N} f(x)_i\, d'\ w_i^{Recency} w_i^{Amplification}\right) + \sigma_{Noise}\right) \quad (4)$$

$$w_i^{Amplification} = e^{A(j/50)} \quad (5)$$

As shown in the above equations, the $w_i^{Amplification}$weight is an exponential function of the item's emotionality, which is denoted by j (1-50), and a free parameter A Note that larger values of A indicate higher weighting of more emotional expressions, and we constrained A to be non-negative (zero inclusive) to capture the fact that there may be no amplification (when A equals zero). Thus, this model has three free parameters, d', r' and A. As before, the Amplification Recency model is like the Baseline and Recency model except that it posits that memory biases are jointly determined by psychophysical similarity, recency effects and amplification of more extreme expressions. Therefore, a comparison of the Amplification Recency model with the Baseline and Recency models provides

[2] This entails that we do not assume that is the best descriptive model of amplification, but rather use it to quantitatively separate amplification biases from psychophysical similarity and recency effects.

direct insight into whether there are amplification memory biases once psychophysical similarity and recency effects are taken into account.

Prior to model fitting, we evaluated each of these models using parameter and model recovery analysis (for discussion of best practices in cognitive modeling see: Heathcote et al., 2015). In our model recovery analysis, we found that Akaike Information Criterion (AIC) is the metric that recovered the true data generating model, therefore, we focus on this metric for the model comparison. We also found that we were able to recover the true data generating model, as long as the presentation of emotional expressions was tightly controlled in the sequence. Study 6 was a good fit for this criteria because the order of the high and low intensity was manipulated, which, based on the model recovery analysis, allows us to differentiate between the competing models. . In studies in which recency is not manipulated, it is hard to differentiate recency from amplification (likely because these effects wash each other out). We fitted these three models to the data using a log-likelihood minimization function and compared their fit using AIC (see Supplementary Table 36). Results suggested that overall, the amplification model yielded the best fit to the data, providing additional support for amplification over and above recency and non-linearity. More specifically, for trials with happy expressions, the recency model was actually the one performing the best out of all models, the second best performing model was the recency amplification model and the last was the baseline model. For trials with angry expressions, the recency amplification model was the strongest fit, with recency second and baseline third. It is important to note that all of the models were fitted to the data of Study 6, in which for positive expressions, no significant amplification was found. These results do not reflect the general trend in many of our other studies and may have been driven by the recency manipulation. Regardless, the fact that the amplification model was the best predictor of the data overall is encouraging. These results provide converging support for the view that amplification is not driven by the psychophysical similarity of our stimulus set, but is indeed driven by memory bias.

Supplementary Table 36. AIC comparisons for three models, the baseline model, the recency model and the amplification model for happy expressions, angry expressions and both.

	AIC		
Model	**Happy**	**Angry**	**Total**
Baseline Model	16275	16561	32835
Recency Model	16119	16422	32540
Recency Amplification Model	16135	16386	32522

7. Study 9: Amplification in the Evaluation of Emotional Videos

Target Videos. Target videos were collected as part of the Stanford Emotional Narratives Dataset (SEND, Ong et al., 2019). Targets were brought into the lab and were asked to think of the three most positive and three most negative events that they would feel comfortable sharing in front of a video camera. Recording was self-paced: The experimenter left the target alone in the room, and targets were allowed to talk for as long as they wanted about each event. After targets finished recording the videos, they were asked to fill out several personality and demographic surveys. During this time, the experimenter processed the videos by transferring them from the camera onto the computer and prepared the next part of the experiment. After targets finished the surveys, they were then shown each video that they recorded. While watching each video, they were asked to provide continuous ratings of how they felt as they were telling their story. These ratings were collected using a visual analog scale divided into a hundred points, ranging from "Very Negative" (-1) to "Very Positive" (+1). The ratings on the scale were sampled every 0.5s. Participants gave their consent to use the videos in future experiments. The subset of video clips selected for the SEND were all consented for research use.

Of the videos that were produced by participants, 193 were selected containing 49 unique targets (Gender: men=20, women=27, other: 2; Age: *M* = 24.8, *SD* = 9.6, ethnicity: East Asian =6, South Asian = 3, Black = 2, Hispanic = 4, Middle Eastern = 1, White = 16, Mixed = 13, Other =4) . This set was chosen such that: (i) the target's face was always in the camera, (ii) the clips did not contain sensitive content (e.g. mental health, suicide), and (iii) the clips were emotional, and had some narrative flow (rather than stream of consciousness or rambling). The clips were also cropped for length, such that the final clips lasted on average 2 minutes 15 seconds. Videos were divided into 4 valence categories by the original authors, which we retained in this study. After transforming video ratings to be on a 0-100 scale (0-very negative, 50-neutral, 50- very positive) videos were divided to four categories. Positive videos included videos that were rated by targets on average as higher than 60, with a minimum rating of .40 (n=62). Negative videos' average were lower than 40 with a max rating of 60 (n=33). Neutral videos were videos that had a max rating of 60 and min rating of 40 (n=30). All other videos were categorized as mixed (68). See original paper for full description of the videos [16].

Participants. We use the terms observers to describe participants who were collected separately at a later date and were asked to provide their evaluation of the target's emotionality. Observers were recruited as part of the SEND database on Amazon Mechanical Turk to watch either videos clips and provide ratings of how the target in the video felt (for details, see Ong et al., 2019). Observers saw each video along with a continuous sliding scale underneath that was designed for continuous emotional ratings. They were asked to dynamically adjust the scale as the video played to capture the emotional intensity of the target at each time point. The analog scale was divided into 100 points (0-very negative, 50-neutral, 100-very positive) and sampled every 0.5s. Severn hundred participants were recruited with the goal of getting at least 20 participants rating each video. Each participant watched 8 videos. The final recruited sample was 695 participants, with 11 additional participants being removed for failing to correctly answer two comprehension checks. Of the remaining 684 participants, we divided the continuous data to windows of 2 seconds and removed any observer ratings for videos that included less than 5 ratings. This elimination standard was different from that of the original researchers who only removed participants who provided zero continuous ratings. We believe that such criteria are a more conservative comparison for the analysis. However, using the original authors' criteria does not change the significance of the results. Our final sample therefore was N=565 (age: *M*=37.23, *SD* =11.23, *gender: female =279, male =254, undefined = 32*).

Methods

Data Reduction. One concern that may be raised when comparing the continuous and post-rating measures is that the continuous rating included the beginning of the videos in which participants did not change their ratings, which meant that their rating was de-facto neutral. Keeping these ratings may artificially reduce the overall average of the continuous rating and further emphasize the amplification. To avoid this issue, we cut the continuous ratings to start only when observers made their first change to the rating, thus removing sections in which the rating was neutral. We then averaged each continuous rating from the point in which participants made their first rating to the end of the video.

Measures. Observers provided two types of ratings in response to each video. The first rating was a continuous rating on a 0-100 scale, 0 indicating very negative, 50 indicating neutral, and 100 indicating very positive. Ratings were sampled every 0.5 second. After watching the video, participants were asked to rate the degree of the target positivity and negativity using two ratings on a 1-7 scale (1-neutral, 7-very emotional), one for positive emotion and one for negative emotions. Because the correlation between positive ratings and negative ratings was very strong (*r* = -.79 [-.75, -.82], and in order to compare the continuous ratings to the post-ratings, we averaged between the positive and negative ratings post-ratings, creating one scale for post-ratings, 1-very negative to 7-very positive. To compare the post-ratings with the continuous ratings we converted the continuous rating to a 1-7 scale by dividing it by 100, multiplying by 6 and adding 1. With this transformation, 100 on a continuous scale was equal to 7 and 0 was equal to 1.

Results

One of the challenges of the current analysis is that amplification may be driven by the fact that continuous and post-measures were evaluated with different scales. In order to account for differences that may have been caused by different scales, we treated the difference between post-rating and continuous rating of the neutral videos as our baseline comparison. Although using neutral videos as a comparison may not solve issues that stem from differences in nonlinearity of the two scales, it can account for baseline differences.

In order to evaluate amplification, we created a difference score between participants' post-rating and their continuous rating such that a positive value indicates that the post-rating was more positive and a negative value indicates that the post-rating was more negative. We then conducted a mixed model analysis using the difference score as our dependent variable and the valence of the video as the dependent variable. Similar to our previous models, we also included by-participant and by-video random intercepts. As previously mentioned, we set our model to use the neutral videos as the baseline comparison (the intercept of the model).

Results suggested that for the neutral videos, the intercept of the model, the difference between participants' post-ratings and their continuous rating was not different from zero (b = 0.16, t (191) = 1.39, p = .17, R^2= .52, 95% Confidence Intervals = [-0.06, 0.39]) providing another indication that neutral condition is a reasonable baseline for comparison. Looking at the negative videos, the difference between the post-rating and the continuous ratings was significantly more negative compared to the difference in the neutral condition (b = -0.92, t (189) = -5.66, p < .001, R^2= .52, 95% Confidence Intervals = [-1.24, -0.60]). On the other hand, and also congruent with the tendency for amplification, the difference between post-ratings and continuous ratings in the positive videos was significantly more positive than the neutral videos (b = 0.70, t (189) = 4.87, p < .001, R^2= .52, 95% Confidence Intervals = [0.42, 0.98], Figure 8).

To further compare the difference between post-ratings and continuous ratings in the positive and negative conditions we multiplied the difference score between the post-ratings and the continuous ratings by -1, thus allowing us to compare the difference score in the positive and negative ratings. We then conducted a mixed model analysis similar to the one above, this time centering the model on the negative videos. Results suggested that amplification in the negative videos was not significantly different in absolute magnitude than that in the positive videos (b = 0.10, t (189) = 0.73, p = .46, R^2= .52, 95% Confidence Intervals = [-0.16, 0.37]).

8. Formal Tests of Assumptions for each Statistical Test Reported in Main Manuscript

Normality

Supplementary Table 37. Kolmogorov-Smirnov test for normal distributions of model residuals. Tests were done for the main analysis of each study (described in the comparison column).

Study	Comparison	D	p
1	Estimated versus actual sequence means.	0.39	< .001

2	Estimated versus actual sequence means.	0.37	< .001
3	Estimated versus actual sequence means.	0.38	< .001
4	Estimated versus actual sequence means.	0.37	< .001
5	Sequence intensity predicts memory	Binomial Distributed	-
6	Impact of recency	0.38	< .001
7	Impact of salience (between groups)	0.39	< .001
9	Comparison between video types	0.079	< .001

Equal variance

Supplementary Table 38. Levene test for equal variance of model residuals. Tests were done for the main analysis of each study (described in the comparison column).

Study	**Comparison**	**F (df)**	**p**
1	Estimated versus actual sequence means.	235.48 (1)	< .001
2	Estimated versus actual sequence means.	67.76 (1)	< .001
3	Estimated versus actual sequence means.	285.52 (1)	< .001
4	Estimated versus actual sequence means.	261.6 (1)	< .001
5	Sequence intensity predicts memory	1.41 (49)	.031
6	Impact of recency	1.94 (1)	.16
7	Impact of salience (between groups)	6.76 (2)	.001
9	Comparison between video types	25.4 (3)	< .001

Robust Estimation vs. Original Estimation

Supplementary Table 39. Robust Estimation of Linear Mixed-Effects Models using *robustlmm*[17]. This method does not make any assumption on the data's grouping structure and is robust against outliers or other contamination.

Study	**Comparison**	**reported b**	**robust b**
1	Estimated versus actual sequence means.	0.75	1.07
2	Estimated versus actual sequence means.	1.37	1.54
3	Estimated versus actual sequence means.	1.32	1.66
4	Estimated versus actual sequence means.	3.96	3.95
5	Sequence intensity predicts memory	0.042	0.010
6	Impact of recency	4.19	4.24
7	Impact of salience (between groups)	2.87	2.86
9	Comparison between video types	0.16	0.18

References

1. Langner, O. *et al.* Presentation and validation of the radboud faces database. *Cogn. Emot.* **24**, 1377–1388 (2010).

2. Goldenberg, A., Weisz, E., Sweeny, T., Cikara, M. & Gross, J. J. The crowd emotion amplification effect. *Psychol. Sci.* **32**, 437–450 (2021).

3. Mattick, R. P. & Clarke, J. C. Development and validation of measures of social phobia scrutiny fear and social interaction anxiety. *Behav. Res. Ther.* **36**, 455–470 (1998).

4. Watson, D., Clark, L. A. & Tellegen, A. Development and validation of brief measures of positive and negative affect: The PANAS scales. *J. Pers. Soc. Psychol.* **54**, 1063–1070 (1988).

5. Gosling, S. D., Rentfrow, P. J. & Swann, W. B. A very brief measure of the Big-Five personality domains. *J. Res. Pers.* **37**, 504–528 (2003).

6. Hughes, M. E., Waite, L. J., Hawkley, L. C. & Cacioppo, J. T. A short scale for measuring loneliness in large surveys: Results from two population-based studies. *Res. Aging* **26**, 655–672 (2004).

7. Barton, K. Multi-Model Inference. (2022).

8. Nakagawa, S., Johnson, P. C. D. & Schielzeth, H. The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *J. R. Soc. Interface* **14**, (2017).

9. Green, P. & Macleod, C. J. SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods Ecol. Evol.* **7**, 493–498 (2016).

10. Hubert-Wallander, B. & Boynton, G. M. Not all summary statistics are made equal: Evidence from extracting summaries across time. *J. Vis.* **15**, 5–5 (2015).

11. Schurgin, M. W., Wixted, J. T. & Brady, T. F. Psychophysical scaling reveals a unified theory of visual memory strength. *Nat. Hum. Behav.* **4**, 1156–1172 (2020).

12. R Core Team. R: A language and environment for statistical computing. (2013).

13. Wickens, T. D. *Elementary signal detection theory*. (Oxford University Press, 2001).

14. Tong, K., Dubé, C. & Sekuler, R. What makes a prototype a prototype? Averaging visual features in a sequence. *Attention, Perception, Psychophys.* **81**, 1962–1978 (2019).

15. Heathcote, A., Brown, S. D. & Wagenmakers, E. J. An introduction to good practices in cognitive modeling. in *An introduction to model-based cognitive neuroscience* 25–48 (Springer Scince, 2015).

16. Ong, D. C. *et al.* Modeling emotion in complex stories: The Stanford Emotional Narratives Dataset. in *IEEE Transactions on Affective Computing* (IEEE, 2019). doi:10.1109/taffc.2019.2955949.

17. Koller, M. robustlmm: An R package for robust estimation of linear mixed-effects models. *J. Stat. Softw.* **75**, 1–24 (2016).