Supporting Information

Thornton and Tamir 10.1073/pnas.1616056114

SI Text

Power Analysis. A Monte Carlo simulation power analysis was used to determine appropriate participant numbers for rating data across studies 1–3. This power analysis targeted the most central planned hypothesis test: whether the average correlation between each participant's transitions ratings and the group-level experienced transitions was greater than 0. On each iteration of the simulation, bivariate normal $N(0, 1)$ data were generated based on a given population-level effect size (rs = 0.1, 0.3), corresponding to small and medium correlations between rated and experienced transitional probabilities. To be maximally conservative, data were simulated with respect to the study with the smallest number of states (18, in study 3) yielding 324 observations of each variable. Simulated data for individual participants were generated by adding additional random normal noise (independent and identically distributed) to one of these vectors for Ns ranging from 20 to 100 in increments of 5. The SD of the added noise was set to produce mean interrater reliabilities of 0.1 and 0.3. Each simulated participant's data were converted to a discrete uniform distribution $U(0, 100)$ to approximate actual ratings, and then Spearman-correlated with the simulated transitions odds from the earlier bivariate normal distribution. Resulting correlation coefficients were linearized using Fisher's r-to-z transformation and entered into a one-sample t test to determine whether they were significantly greater than 0 at the $\alpha = 0.05$ level. We used t tests in the power analysis, rather than the bootstrapping adopted with the actual data, because of the computational efficiency of the former relative to the latter. This process was repeated for 5,000 iterations at each combination of effect size, sample size, interrater reliability.

The pooled results indicated that the design in question was intrinsically quite well-powered, because of the number of ratings each participant provided. With as few as 20 participants (the smallest sample considered), a combination of moderate effect size (0.3) and moderate interrater correlation (0.3) guaranteed ∼100% power. As expected, decreasing either the effect size or the interrater reliability diminished power, with a superadditive effect of decreasing both simultaneously. The simulation indicated that with an effect size of $r = 0.1$ and a mean interrater of $r = 0.3$, 30 participants would provide 80% power. An effect size of $r = 0.3$ with a mean interrater of $r = 0.1$ would require 35 participants to provide 80% power. If both the effect size and reliability were small (0.1), then increasing power would be highly inefficient: 35 participants would provide 32% power, but even 100 participants would only yield 47% power. These results suggested that a sample size as low as $n = 35$ would provide acceptable power, but to improve our estimates of effect sizes we targeted a sample twofold larger: $n = 70$, after exclusions. This sample size was estimated to provide nearly 100% power if both the interrater correlation and the accuracy effect size were moderate (0.3), 98% power if the interrater correlation was small and the effect size was moderate, 88% power if the interrater correlation was moderate and the effect size was small, and 40% power if both were small. Moreover, these estimates were based on testing only 18 states, consistent with study 3 but conservative for studies 1 and 2.

An additional simulation-based power analysis was conducted for study 5. This power analysis was targeted at one of the key goals of this study: determining whether transitional probability ratings had incremental validity in predicting ground-truth positions over and above similarity ratings. The simulation paralleled the actual analysis, although simplifying measures were taken for computational tractability: t tests rather than bootstrapping were used to assess statistical significance, simulated responses were drawn from a normal distribution, and Pearson rather than Spearman correlations were used. We fixed the correlation between (group-averaged) simulated transition ratings and simulated ground-truth transitional probabilities to $r = 0.7$, based on the results of studies 1–3. To be highly conservative with respect to the incremental value of transition ratings, we fixed the correlation between similarity ratings and ground-truth transitional probabilities to $r = 0.69$, and the correlation between similarity ratings and transition ratings to 0.99. The reliabilities from studies 1–3 were used to determine the signal-to-noise ratio for adding random variance to simulated individual participants. On each iteration of the simulation, we calculated whether the average partial correlation between individual participant transition ratings and ground truth (accounting for group-averaged similarity ratings) was greater than 0. We found that a sample size of 150 would be sufficient to provide 97% power at $\alpha = 0.05$, even under these conservative assumptions.

Manipulation Check. A covert attention check was included in the rating paradigm for studies 1–4: a radio button item in the middle of the demographic posttest, which appeared to ask about United States nationality, but actually instructed participants not to respond. Although it was initially our intention to use this as an exclusion criterion, this check proved more difficult for participants than we anticipated, eliciting very high failure rates (42%, 32%, and 49% of otherwise included participants in studies 1–3). Furthermore, we found that the average interrater reliabilities were nearly numerically identical regardless of whether participants who completed this attention check were excluded, suggesting that this check was not related to data quality. Note that this reliability check is not directly related to our hypothesis test, and thus does not bias results. As a result of these indications of the dubious validity of the manipulation check, we retained all participants in the analyzed samples. We subsequently dropped this check in study 5.

Frequency-Normalizing Experienced Transition Matrices. Raw transitional probability matrices, derived from the experience-sampling data in studies 1–3, were normalized by frequency expectations. Frequency-based expectations for each cell in the transition matrix were calculated by summing the occurrences of each emotion and then multiplying the resulting vector by its transpose. The resulting matrix was divided by its sum and then divided into the transitional probability matrix (also sum-normalized) elementwise.

Frequency normalization of the experienced transitions was undertaken for two reasons. First, without such normalization, the frequencies would likely dominate other sources of variance in the experienced transition odds matrix. This would have make it difficult to determine whether any observed accuracy resulted from meaningful mental models of emotional transitions per se, or just knowledge of the frequencies. Second, the common tendency toward base-rate neglect led us to anticipate that participants would make relative judgments of transition likelihoods that ignored the global frequencies of each emotion. Failing to normalize for frequencies would thus have unnecessarily contaminated accuracy estimates with a known source of cognitive bias.

Experience Project Transitional Probabilities. The transitional probabilities that were provided to us were calculated in earlier work (21) by applying an exponential decay model to 2 million mood reports on the website. Thus, all states subsequent to a given emotion were considered, but down-weighted exponentially based on their temporal distance to the emotion report in question. The coefficient of this exponential model was set such that reports "t" days after the

initial report would matter half as much as reports "$t - 1$" days after that report. Thus, the transitional probabilities in study 5 incorporate the time intervals between reports in a relatively naturalistic manner.

The transitional probabilities we were provided were normalized with respect to the state first experienced, but not with respect to the endpoint of the transition. Given that the transitional probabilities were approximately distributed according to a power law (and thus scale-free), we subtracted 1% from these values, took the mean of the values between 0 and 1%, and imputed this mean to undefined cells of the transition matrix. We then normalized with respect to the transition sums of the endpoint emotions. We included only pairs of emotions with bidirectional transitions available, leaving a final set of 456 ground-truth transitions over 57 states.

Frequency Analyses. Participants' ratings of their own emotional frequencies in studies 1–3 were subjected to analogous consensus and accuracy analyses as those conducted with the transitional probability ratings. The average correlations among participants for the frequency ratings were rs = 0.19, 0.25, and 0.38, respectively. In each case, the frequency consensus was less than the transitional probability consensus, suggesting that participants had more homogeneous intuitive models than actual emotional experiences. Item analysis of the intuition and experience sampling emotional frequencies yielded correlations of ρs = 0.68, 0.70, and 0.75. Individual participant ratings correlated with the group-level experience ρs = 0.31, 0.38, and 0.48, all of which were significantly significant at $P < 0.05$ as assessed by percentile bootstrapping.

Out-of-Sample Prediction. To complement the primary inferential statistical approach in the paper—which relied on nonparametric null-hypothesis significance testing—we also conducted cross-validated prediction with respect to the primary accuracy relations in each study. Fivefold cross-validation was used in each case. In studies 1–3, we cross-validated with respect to both participants and emotions. In the former case, we calculated separate simple regressions predicting each participant in the training set's responses with transition log odds from the corresponding experience sampling study. Using the average regression parameters (slope and intercept) from these regressions, we then predicted the transition ratings of participants in the left-out test set. This process was repeated iteratively leaving out each of five randomly divided "folds" in the sample. In the latter case (i.e., cross-validating with respect to emotions), on each fold we fit a simple regression predicting group-average transition ratings from experienced transition log odds for the emotions in the training set. We then predicted group-average transition ratings for the left-out emotions using the model fit to the training set. Prediction error for both the participant and emotion cross-validation was calculated using the formula for RMSE. Also, in both cases we computed RMSE for a test set that had been randomly permuted with respect to emotion pairs.

In study 4, we performed a simpler cross-validation of the correlation between self-reported frequencies and rated transition matrix stationary distribution. In each case, we fit a linear regression to this relationship using four-fifths of the 60 states, and then predicted the self-reported frequencies of the left out set using the corresponding stationary distribution values and the pretrained model. In study 5, we perform cross-validation with respect to participant in the manner described for studies 1–3, but did not also perform emotion cross-validation because of the sparse nature of the ground-truth transitional probability matrix in this dataset.

Results across all five studies supported the conclusion that participants' mental models of emotion transitions were indeed accurate. In studies 1–3, the RMSEs in the participant-wise cross-validation were 28.49, 27.36, and 27.58. The corresponding RMSE values with randomized permuted test sets were 34.02, 30.62, and 34.51. RMSEs in the emotion-wise cross-validation were 18.13, 12.06, and 17.90, with RMSEs in the randomized equivalents of 24.83, 16.62, and 21.93. The RMSE of the correlation in study 4 was 14.17, with a randomized baseline of 21.41. The participant-wise RMSE in study 5 was 28.92, and the corresponding randomized RMSE was 29.98. Range-normalized versions of these values are reported in the results section. Note that in every case RMSE was higher for the randomly permuted test sets then for the properly order sets in the same analysis, indicating the above-chance performance of the participants' predictions.

Dimensional Mediation Analysis. In study 5, we tested whether the four conceptual dimensions tested in study 4 mediated the relationship between transition ratings and ground truth in study 5. Two of the three legs of this analysis were completed as described in the main text: correlating dimension ratings with transitional probability ratings, and correlating dimension ratings with ground-truth transitions. The third leg of the mediation examined whether the accuracy relationship between transition ratings and ground truth changed as a function of controlling for the dimensions. In this analysis, we calculated the partial correlation between individual participant's transitions ratings and ground-truth transitional probabilities, controlling for aggregate dimension ratings. We then recalculated the latter value leaving out each dimension in turn and assessing the change in partial correlation in each case. Large increases in accuracy partial correlation as a function of leaving out a dimension (in conjunction with that dimension correlating with both ratings and ground truth) was taken as indicative of statistical mediation of the accuracy. Statistical significance in all portions of the mediation analysis was calculated via bootstrapping. We observed significant increases in the residual accuracy relationship when removing valence, social impact, and rationality from the model, but not human mind [mean change in partial ρs = 0.05, 0.03, 0.01, –0.0008, 95% percentile bootstrap CIs = (0.049, 0.056), (0.028, 0.033), (0.010, 0.013), (–0.001, –0.0005)]. Together with the fact that these dimensions are associated with both transition ratings and ground truth, these results provide evidence that valence, social impact, and rationality each uniquely mediate part of the accuracy of people's mental models of emotion transitions.

The Egocentricity of Mental Models. One possible source of inaccuracy is egocentric bias: participants' own idiosyncratic emotional experiences may influence their intuition about others'. We took a discriminative approach to assessing the effect of egocentrism on participants' ratings of transitional probabilities. We correlated participants' frequency ratings with each other, and did the same with participants' transitional probability ratings. We linearized both correlation matrices using Fisher's r-to-z transformation, and then correlated the lower triangular portion of these two correlation matrices. The statistical significance of this relationship was assessed by permutation testing. In this case, the rows and columns of the two correlation matrices were permuted, thus treating the participant as the level of independent observation. If participants' idiosyncratic experiences egocentrically biased their mental models, then participants with similar emotional experiences, as assessed by their frequency reports, should also have similar models. We observed a small but reliable impact of idiosyncratic emotion experiences on mental models, with significant correlations between frequency- and transition-similarity matrices (rs = 0.14, 0.17, 0.18; Ps = 0.0495, 0.016, 0.018) in studies 1–3. This relationship suggests that participants may partially base their models of others' emotion transitions on their own emotion transition experiences.

Co-Occurrence Analysis. For studies 1–3 we calculated emotion co-occurrence matrices using the same method used to calculate transitional probability log odds. In each case, we observed large correlations between these co-occurrence matrices and the corresponding transitional probability matrices (ρs = 0.97, 0.90, 0.99). We also observed high correlations between the co-occurrence

matrices and group-average transition ratings (ρs = 0.82, 0.72, 0.82). The association between transition ratings and ground truth appeared to be fully mediated by the variance these variables shared with the co-occurrence matrices, as the average partial correlations were not significantly greater than 0 when controlling for co-occurrence odds. These results suggest that people may take advantage of the very high ecological correlation between co-occurrences and transitional probabilities by using their knowledge of the former (which is easier to acquire, requiring only a single observation) to inform their judgments of the latter. However, it should be noted that there were considerable gaps between the proportions of true variance in the group-average mental models (i.e., their reliabilities), and proportion of variance explained in these models by the co-occurrences. Thus, the transition models are not themselves completely explained by co-occurrence. Indeed, it is mathematically impossible for the co-occurrences to explain certain features of the mental models, such as the robust asymmetries in transitional probabilities for particular pairs of states or variability along the diagonal rating transitional probability matrix.

Analysis of Residuals in the Frequency–Stationary Distribution Relationship. A substantial correlation between the emotion frequencies participants self-reported in study 4 and the stationary distribution calculated from their transitions ratings, but it is worth considering whether any identifiable factors account for the errors in this model. To this end, we recalculated the correlation as a simple regression with self-reported frequencies as the dependent variable. We then calculated the correlations between these residuals and the four dimensions of mental state representation we consider in studies 4 and 5. We found that rationality ($r = 0.36$) and valence ($r = 0.47$) were positively correlated with these residuals, suggesting that the stationary distribution overpredicted the self-reported frequencies of affective (vs. cognitive) states and negative states. The dimension of human mind was negatively correlated with residuals ($r = –0.26$), suggesting that uniquely human mental states were also overpredicted. Social impact expressed a very small correlation ($r = 0.03$) with residuals, indicating that this dimension was not related to the accuracy of the stationary distribution.

Exponential Decay Modeling. Following ref. 21, we fit exponential decay models to the experience-sampling data in studies 1–2 to explore the characteristic time-scale of the emotions under investigation. These data were particular well-suited to this analysis, as the experience-sampling was relatively frequent (every 3 h), and there were large numbers of within-participant reports (>70), in comparison with the experience-sampling data in study 3. Within participant, and for each state in each study, we calculated the time between an emotion being reported and all subsequent experience-samples, while also calculating whether the emotion was present at the time of those subsequent reports. We then fit an exponential decay model consisting of a binary logistic regression predicting whether the emotion was present absent at the subsequent time-point. The single predictor in this model was the natural logarithm of the time difference between pairs of reports.

All but one emotion was best fit by a negative coefficient, indicative of decay in the probability of recurrence of an emotion over time (Fig. S4*A*). The single exception to this rule was the state of "calm" in study 1. This exception might occur because participants viewed clam as a neutral baseline state to which they would return by default, thus leading on an increasing probability of recurrence over time. The otherwise universal decay of states suggests that the experience under study cannot be reduced to trait or dispositional tendencies to report certain states.

Using the logistic regression fits, we were able to calculate the emotional "half-life" of each state. This value corresponded to the time it took for the emotion to decay 50% of the time on average, correcting for its baseline frequency. A one-dimensional optimization procedure probed the fitted models to find the point that yielded the appropriate recurrence rates. The resulting half-lives were almost all less than 1 h, with most in the range of 5 min or less (Fig. S4*B*). Given that the experience-sampling rate was only once every 3 h, these estimates are naturally extrapolations from the observed data. However, these extrapolations rely only on the assumption that emotions undergo exponential decay, which is minimal, plausible, and has precedent in the literature (21). The short characteristic time-scales of these states suggests that they are indeed emotions in the typical sense, rather than more temporally extended moods. This result also suggests that very high density or continuous experience-sampling would likely yield much high signal-to-noise in studying emotion transitions. The results observed in the present investigation rely solely on the long tails of the exponential decay distributions for observable signal, whereas much of the meaningful variance occurs on a shorter time-scale.

Full Transition Rating Task Instructions. Below we reproduce the full instructions presented to each participant at the beginning of the transition rating task. In this case the instructions were taken from study 5, but very similar instructions were used for each experiment.

> People can experience many different emotions. These emotional states are not static. Instead, they change gradually over time. We are interested in your thoughts on how one emotion may lead to another. So for example, if a person feels tired one moment, what are the chances that they will next feel excited? Or what are the chances that they will next feel sleepy instead?
>
> In this study you will be presented with pairs of emotions. The first emotion denotes a person's current state; the second emotion denotes an emotional state that person could potentially feel next. You task is to estimate the likelihood of a person currently feeling the first emotion subsequently feeling the second emotion. For this example, what is the chance of a person currently feeling tired next feeling excited?
>
> For example, this transition will be presented as:
>
> Tired → Excited
>
> You will make your rating on a scale from 0 to 100%, where 0% means that there is zero chance that a person feeling tired will feel excited next, and where 100% means that a person feeling tired now will definitely feel excited next.

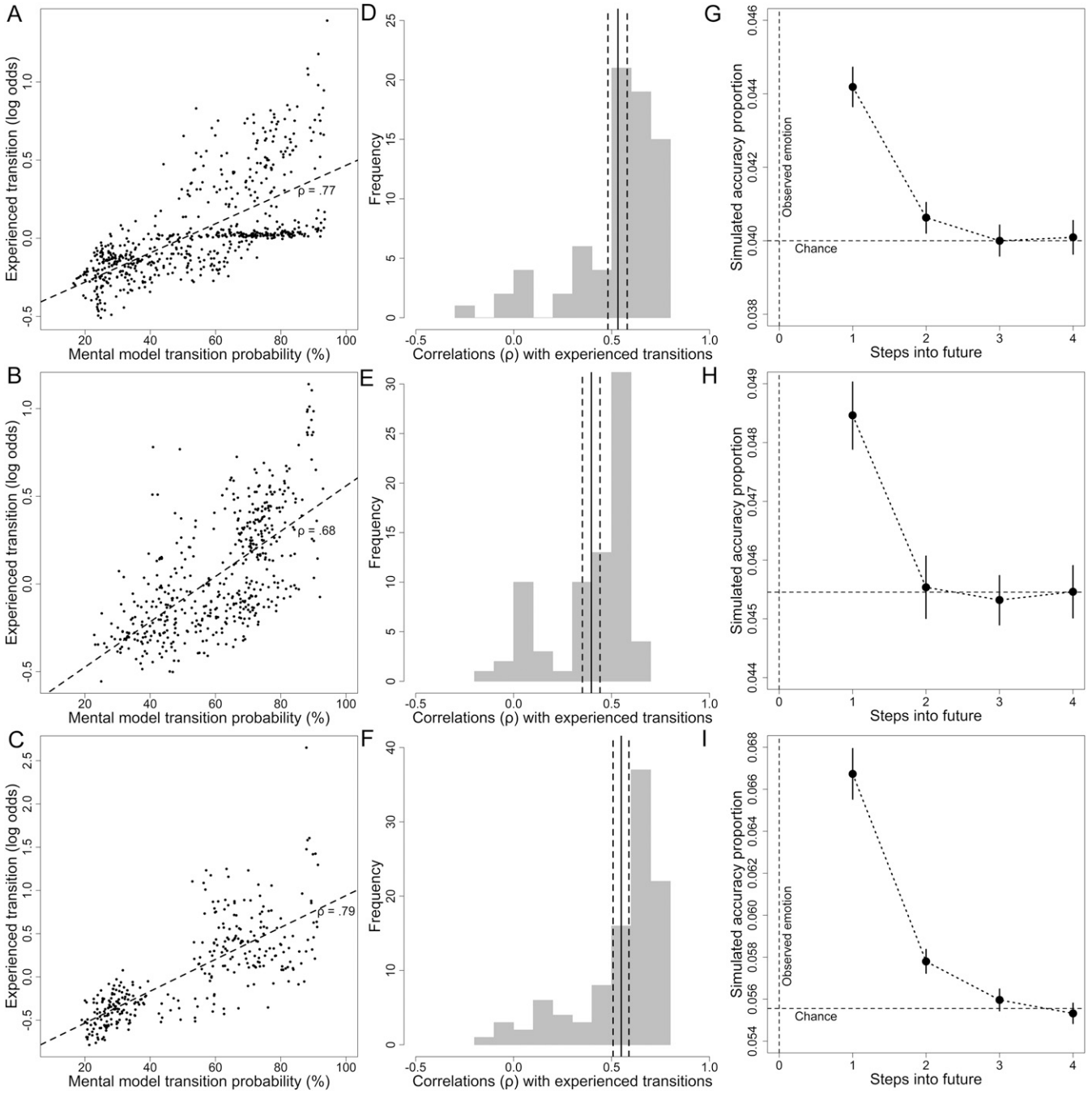

Fig. S1. Mental models accurately predict actual emotion transitions in studies 1–3. (*A*–*C*) The relationship between the log odds of transitions from three experience-sampling datasets and corresponding mental models of emotion transitions, averaged across distributions of accuracy of individual participants' models of emotion transitions from three datasets. Each point corresponds to the transition likelihood between a pair of emotions; dashed lines indicate linear best fit. (*D*–*F*) The distributions of accuracy of individual participants' models of emotion transitions from three datasets. Solid vertical lines indicate the mean correlation coefficient, and dashed lines indicate 95% CI calculated via percentile bootstrapping. (*G*–*I*) The accuracy of individual participants' mental models at each step in a random walk through experience-sampled emotion transition matrices. Horizontal dashed lines indicate accuracy expected from random guessing. Error bars indicate 95% CIs calculated via percentile bootstrapping.

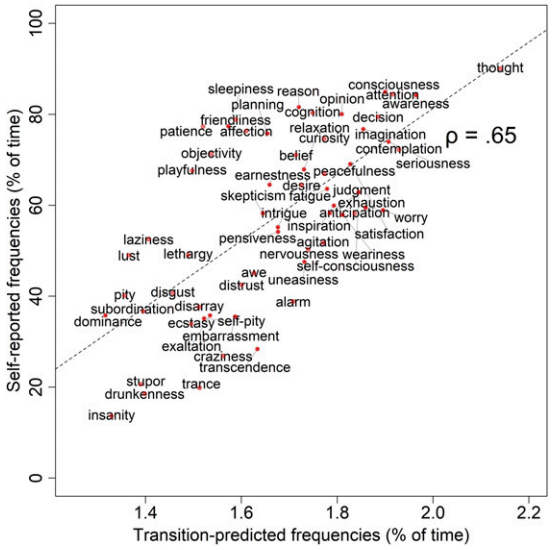

Fig. S2. Accurate frequency predictions from mental models in study 4. The high correlation between rated mental-state frequencies and the stationary distribution of the mental model Markov chain provides convergent evidence for the accuracy of people's mental models.

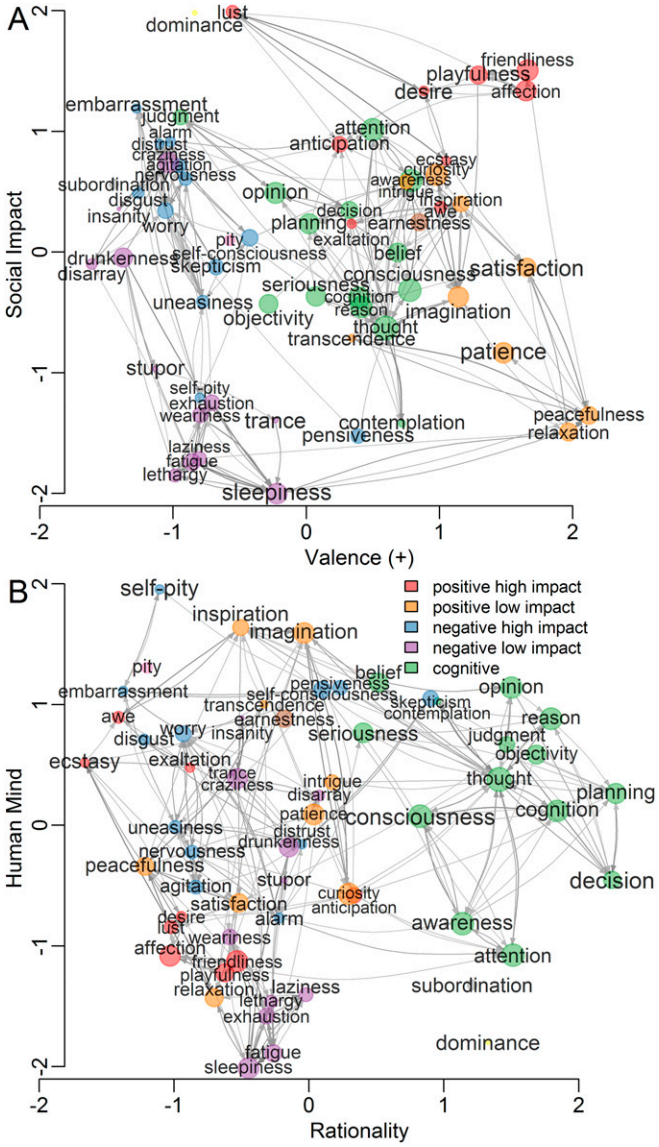

Fig. S3. Emotion transition networks in a 4D representational space. The network graphs represent likely transitions (>75%) between mental states in study 4. Node colors indicate optimal (modularity maximizing) clusters of states that transition to each other. Descriptive labels for multinode clusters are provided for convenience. Node size indicates how frequently participants experienced the state. The positions of the states reflect where they fall on the psychological dimensions of valence and social impact (*A*) or rationality and human mind (*B*). The effects of the dimensions of rated transitions can be observed in the relatively sparsity of long-distance links in comparison with short-range links. These effects are also reflected in the spatial clustering of nodes of the same color cluster in the 4D representational space.

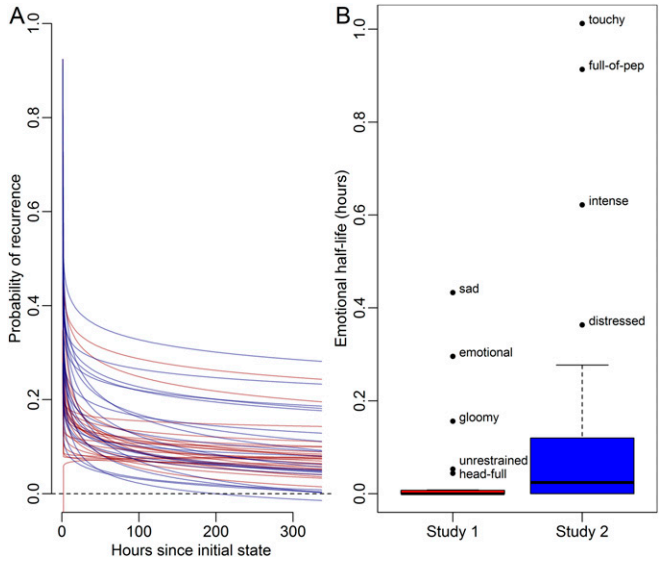

Fig. S4. Exponential decay and half-lives of emotions. The figures represent statistics of the exponential decay of emotions in the experience-sampling data from studies 1 (red) and 2 (blue). In *A*, the characteristic decay curves from binary logistic regressions are plotted for each of the emotions in these two studies. Overall frequencies were subtracted from the curves to adjust for the different base rates of each emotion. For all but one emotion (calm) the fitted models indicate decay: that is, decreasing probability of recurrence with time. In *B*, the boxplots illustrate the distribution of emotional "half-lives" derived from these exponential decay models. The half-life is the time it takes for the probability of recurrence to drop below 50%, again correcting for base-rates by subtraction. Outlier emotions with long half-lives are individually labeled.

Table S1. Demographic breakdown and exclusions for samples

Study	Rating	Total sample size (*n*)	Language exclusions	Unique response exclusions	Final sample size (*n*)	Female (*n*)	Male (*n*)	Mean age (y)	Minimum age (y)	Maximum age (y)
1	Transitions	80	6	0	74	38	36	35.4	20	66
2	Transitions	82	6	0	76	51	25	32.1	19	59
3	Transitions	109	6	1	102	53	49	34.5	19	69
4	Transitions	337	32	3	302	183	119	38.0	19	73
5	Transitions	152	1	0	151	69	81	36.6	20	70
5	Similarity	154	1	4	149	70	79	36.4	19	70
5	Rationality	44	1	1	42	21	21	38.4	22	66
5	Social impact	46	1	3	42	24	18	38.7	21	67
5	Valence	49	2	2	45	19	26	37.0	21	60
5	Human mind	47	0	4	43	23	20	37.4	22	67