Emotion

(Not) Hearing Happiness: Predicting Fluctuations in Happy Mood From Acoustic Cues Using Machine Learning

Aaron C. Weidman, Jessie Sun, Simine Vazire, Jordi Quoidbach, Lyle H. Ungar, and Elizabeth W. Dunn

Online First Publication, February 11, 2019. http://dx.doi.org/10.1037/emo0000571

American Psychological Association

Emotion

1528-3542/19/$12.00

http://dx.doi.org/10.1037/emo0000571

(Not) Hearing Happiness: Predicting Fluctuations in Happy Mood From Acoustic Cues Using Machine Learning

Aaron C. Weidman
University of Michigan

Jessie Sun and Simine Vazire
University of California, Davis

Jordi Quoidbach
ESADE Business School

Lyle H. Ungar
University of Pennsylvania

Elizabeth W. Dunn
University of British Columbia

Recent popular claims surrounding virtual assistants suggest that computers will soon be able to hear our emotions. Supporting this possibility, promising work has harnessed big data and emergent technologies to automatically predict stable levels of one specific emotion, happiness, at the *community* (e.g., counties) and *trait* (i.e., people) levels. Furthermore, research in affective science has shown that nonverbal vocal bursts (e.g., sighs, gasps) and specific acoustic features (e.g., pitch, energy) can differentiate between distinct emotions (e.g., anger, happiness) and that machine-learning algorithms can detect these differences. Yet, to our knowledge, no work has tested whether computers can automatically detect *normal, everyday, within-person fluctuations in one emotional state* from acoustic analysis. To address this issue in the context of happy mood, across 3 studies (total $N = 20{,}197$), we asked participants to repeatedly report their state happy mood and to provide audio recordings—including both direct speech and ambient sounds—from which we extracted acoustic features. Using three different machine learning algorithms (neural networks, random forests, and support vector machines) and two sets of acoustic features, we found that acoustic features yielded minimal predictive insight into happy mood above chance. Neither multilevel modeling analyses nor human coders provided additional insight into state happy mood. These findings suggest that it is not yet possible to automatically assess fluctuations in one emotional state (i.e., happy mood) from acoustic analysis, pointing to a critical future direction for affective scientists interested in acoustic analysis of emotion and automated emotion detection.

Keywords: happiness, happy mood, experience-sampling, acoustic analysis, machine learning

Supplemental materials: http://dx.doi.org/10.1037/emo0000571.supp

If you have been reading the popular press, you might have the impression that we are nearing a time in which computers will able to automatically hear emotions. Multinational corporations such as Amazon, Google, and Apple are working to roll out voice-based emotion recognition features to accompany virtual assistants such as *Amazon Alexa* (e.g., Knight, 2016; Vogels, 2018). Moreover, technology start-ups currently boast products such as in-car microphones that can detect drivers' emotions (Affectiva, 2018) and virtual assistants that can appraise a user's mood to recommend a certain restaurant or movie (Beyond Verbal, 2018).

Developing automated, computerized methods to detect emotion through acoustic analysis would benefit affective scientists as well. The field is currently experiencing a technological revolution in which researchers aim to capture *life as lived* through smart-

Aaron C. Weidman, Department of Psychology, University of Michigan; Jessie Sun and Simine Vazire, Department of Psychology, University of California, Davis; Jordi Quoidbach, Department of People Management and Organisation, ESADE Business School; Lyle H. Ungar, Department of Computer and Information Science, University of Pennsylvania; Elizabeth W. Dunn, Department of Psychology, University of British Columbia.

Aaron C. Weidman and Elizabeth W. Dunn developed the study concept and performed data collection for Study 1. Jordi Quoidbach performed data collection for Study 2. Jessie Sun and Simine Vazire performed data collection for Study 3. Aaron C. Weidman performed data analysis. Elizabeth W. Dunn, Simine Vazire, Jessie Sun, Jordi Quoidbach, and Lyle H. Ungar provided critical insight into how the manuscript should be framed and how the results should be interpreted. Aaron C. Weidman wrote the manuscript, with assistance from Jessie Sun and Jordi Quoidbach. All authors contributed to editing subsequent versions of the manuscript and approved the final version for submission.

Correspondence concerning this article should be addressed to Aaron C. Weidman, Department of Psychology, University of Michigan, 530 Church Street, Ann Arbor, MI 48105. E-mail: weidman@umich.edu

phones and digital sensing devices (Harari et al., 2016; Mehl & Conner, 2013; Nelson & Allen, 2018). Yet, the current modal method for assessing emotions—self-report—is problematic in that reporting one's emotions can alter those emotions (e.g., Kassam & Mendes, 2013; Lieberman, Inagaki, Tabibnia, & Crockett, 2011) and can introduce distortions when done repeatedly across time and context (e.g., contrast and decline effects; Baird & Lucas, 2011; Schwarz, 1999; Shrout et al., 2018). There are also pragmatic limits to the amount of emotion self-reports that people can complete in a given time period of interest (e.g., 10 reports a day for 2 weeks; Kuppens, Oravecz, & Tuerlinckx, 2010). Furthermore, even in these intensive experience-sampling protocols, participants will at times fail to report their emotion. The implication is that using self-report to assess momentary emotion will at best yield an incomplete portrait of a person's emotion during the time period of interest and at worst will result in some distortions when emotion is reported.

Acoustic analysis provides a particularly promising medium through which to explore the possibility that emotions can be assessed without self-report: Acoustic recordings can easily and unobtrusively be collected via smartphones—which nearly 80% of Americans now own (Pew Research Center, 2018)—and can be collected continuously throughout a time period of interest. Acoustic recordings are, therefore, less susceptible to issues of missing data compared with self-report assessments or other forms of naturalistic data such as social media posts (e.g., a person may only Tweet twice in a given day, making it difficult to track their emotion between these time points). A method to automatically detect emotion from acoustic analysis would, therefore, provide an instant, near-ubiquitous window into people's emotions for research purposes.

To our knowledge, however, there have been no empirical tests of whether normal, everyday fluctuations in any specific emotional state can be detected through acoustic analysis in an automated manner (i.e., without intensive human effort). The purpose of the present article is to test this question. We focused on one particular emotional state—happy mood—because variability in mood (or valence) is widely considered to be a dimension underlying all other emotional experiences (Russell & Barrett, 1999; Watson & Tellegen, 1985).

Leveraging Technology to Automatically Predict Happy Mood: Prior Promising Findings

Two ongoing lines of work speak broadly to the goal of predicting fluctuations in happy mood through acoustic analysis: (a) work showing that big data methodologies other than acoustic analysis can be used to predict stable happiness at the community (e.g., counties) or trait (i.e., people) level; and (b) work showing that acoustic analysis can be used to differentiate between distinct state (i.e., momentary) emotional experiences. We review both of these lines of work below.

Predicting Happiness at the Community or Trait Level

Recent technological advances in big data methodologies other than acoustic analysis have allowed psychologists to automatically predict stable levels of community and trait happiness. At the *community* level, Dodds and colleagues (2011) attempted to track the aggregate happiness of the United States by indexing the positivity of words used in 26.5 billion American Tweets; however, this study did not connect automated happiness ratings to a self-report criterion. Schwartz and colleagues (2013) overcame this issue by showing that aggregate self-reported happiness among inhabitants of 1,293 American counties correlated moderately ($r = .31$) with the relative positivity of words used in Tweets made by a separate sample of those counties' inhabitants.

At the *trait* level, Kosinski, Stillwell, and Graepel (2013) showed that self-reported happiness of 58,466 Facebook users correlated .17 with automated predictions made through machine learning analysis of these users' Facebook likes. Similarly, Schwartz and colleagues (2014) showed that self-reported depression—which might be considered extremely low happiness—among 28,749 Facebook users correlated .39 with automated predictions made through machine learning analysis of these users' Facebook statuses.

Differentiating Between Distinct State Emotions Via Acoustic Analysis

Affective scientists have long contended that emotions are communicated through the voice (e.g., Scherer, 1986). Supporting this notion, recent work has shown that distinct emotions are conveyed through distinct nonverbal vocal bursts (e.g., Cordaro, Keltner, Tshering, Wangchuk, & Flynn, 2016; Sauter, Eisner, Ekman, & Scott, 2010). For example, Sauter and colleagues (2010) found that vocal bursts are sufficient for emotions such as anger (conveyed with a growl), fear (conveyed with a scream), surprise (conveyed with a sharp inhalation), and amusement (conveyed with a laugh) to be recognized by both English speakers and members of the remote, isolated Himba society. Complementary work has shown that specific acoustic features (e.g., pitch, energy) can help distinguish speech utterances conveying distinct emotions such as anger and happiness (e.g., Banse & Scherer, 1996; Wallbott & Scherer, 1986).

Recent work bridging affective science and computer science has further utilized multivariate machine learning to classify distinct emotional vocalizations based on acoustic cues. For example, Laukka, Neiberg, and Elfenbein (2014) used machine learning analysis of 30 acoustic features to classify speech utterances made by professional actors as reflecting one of 11 distinct emotions at above-chance levels. In a more naturalistic study, Karam and colleagues (2014) used machine learning analysis of 23 acoustic features extracted from phone conversations to accurately classify bipolar disorder patients as experiencing manic or depressed mood—states which are characterized by widely divergent psychological and physiological characteristics that might be expected to manifest in distinct vocal signatures (American Psychiatric Association, 2013).

Our Focus: Predicting Fluctuations in *State* Happy Mood From Acoustic Analysis

The present article focused on a distinct theoretical question than the work reviewed above: Can fluctuations in one particular emotional state, *happy mood*, be predicted from acoustic analysis? This question differentiates our work from prior work predicting happiness at the community or trait level using big data methodologies other than acoustic analysis (e.g., Dodds et al., 2011;

Kosinski et al., 2013; Schwartz et al., 2013, 2014), in that we aim to track moment-to-moment fluctuations in happy mood rather than stable levels of happiness within a community or a person. This question also differentiates our work from prior work distinguishing between distinct state emotions via acoustic analysis of both vocal bursts (e.g., Cordaro et al., 2016; Sauter et al., 2010) and speech utterances (e.g., Banse & Scherer, 1996; Karam et al., 2014; Laukka et al., 2014), in that we aim to differentiate between levels of the same emotion (i.e., high vs. low happy mood) rather than between distinct emotions (e.g., anger and happiness).

Although prior work in this domain gives us reason for optimism that we may be able to predict fluctuations in state happy mood from acoustic analysis, the differences outlined above provide reasons for pause: It may be far more challenging to predict moment-to-moment variability in levels of happy mood, compared to predicting stable community or trait happiness, or compared with differentiating between distinct emotions. This is because the acoustic differences between the same individual experiencing high versus low happy mood may be subtle (i.e., within-person differences) whereas we might expect two individuals who typically experience high versus low happiness to differ in many ways that are detectable through big data analyses (i.e., between-person differences). We might also expect that distinct emotions such as anger and happiness produce widely divergent acoustic signatures, particularly when enacted by professional actors as in prior work (e.g., Laukka et al., 2014).

Nonetheless, determining whether it is possible to predict fluctuations in state happy mood from acoustic analysis would speak to the goal of automatically assessing emotion without self-report. Acoustic cues can be collected and analyzed from smartphone-based recordings in a manner that makes them amenable to automated assessment among smartphone users (which, as noted above, constitutes the majority of the population; Pew Research Center, 2018). In contrast, although the social media data used to predict stable community and trait happiness in prior work is amenable to automated analysis, most people do not Tweet frequently enough to provide sufficient data to use a machine learning model to reliably predict fluctuations in state happy mood. For example, if a person Tweeted at 8:14 a.m. and again at 3:45 p.m., it would be very difficult to predict their state happy mood at 11:00 a.m. or 7:00 p.m. from the content of those two Tweets alone. In light of these issues, as well as the aforementioned limitations in self-reported emotion (e.g., Kassam & Mendes, 2013; Lieberman et al., 2011; Shrout et al., 2018), a method providing automated insight into fluctuations in happy mood from acoustic cues would be extremely useful. Therefore, we aimed to test whether fluctuations in state happy mood can be automatically predicted via acoustic analysis.

The Present Research

We harnessed three independently collected samples to test whether fluctuations in state happy mood can be predicted from acoustic analysis. In each study, we took three steps to maximize our chances of successfully predicting state happy mood. First, we collected self-reported state happy mood repeatedly over several days using intensive, smartphone-based experience-sampling (ESM). Given that ESM is seen as the gold standard for capturing people's in vivo experiences (e.g., Conner, Tennen, Fleeson, & Barrett, 2009), we treated these self-reports of state happy mood as a "ground-truth" criterion in our predictive analyses.

Second, we collected audio recordings from participants that corresponded to the time at which these state happy mood reports were made. For each recording, we extracted a set of acoustic features that was tailor-made to detect emotion: The Extended Geneva Minimalist Acoustic Parameter Set (eGeMAPS; Eyben et al., 2016), which was derived by affective scientists based on prior theoretical and empirical links between acoustic features and emotion (e.g., Banse & Scherer, 1996; Scherer, 1986).

Third, we used three distinct machine learning algorithms—neural networks, random forests, and support vector machines—to predict state happy mood from the acoustic features extracted from each recording. Compared with conventional regression, machine learning algorithms have multiple advantages, including (a) the ability to handle many predictors without overfitting and (b) the ability to fit nuanced models involving complex interactions among predictors and nonlinear links between predictors and outcomes (e.g., Strobl, Malley, & Tutz, 2009). Machine learning algorithms have recently been used to gain insight into difficult prediction-related psychological research questions (e.g., Joel, Eastwick, & Finkel, 2017; Laukka et al., 2014; Wang & Kosinski, 2018).

Across all studies we report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. Data, code, and materials have been made available on the OSF (https://osf.io/4pzf7/).

Study 1

Method

Participants. Five-hundred and seventy-five students from the University of British Columbia enrolled in the study in exchange for course credit. Seventy-three of these participants did not complete the experience-sampling phase, leaving a final sample size of 502 (76% women, M_{age} = 20.55, SD = 2.99, 41% East Asian, 23% European, 11% South Asian, 7% Southeast Asian, 4% Middle Eastern, 14% other). We recruited the maximum number of participants that we could during one academic semester given our university's resources. Procedures for this study received ethical approval from the University of British Columbia Behavioral Research Ethics Board.

Procedure. Participants completed an initial lab assessment during which they reported demographics and trait happiness, using the Subjective Happiness Scale (Lyubomirsky & Lepper, 1999; M = 4.81, SD = 1.08, α = .84). Next, in the experience-sampling portion of the study, participants were sent a survey link via text message during a randomly selected two-hour window between 10:00 a.m. and 8:00 p.m. for six consecutive days. Participants were asked to complete each survey as soon as possible and were told to disregard the survey if they had not completed it by the time they received the next day's survey. Each survey involved three tasks: reporting state happiness and making two audio recordings on a smartphone. These tasks were presented in a randomly determined order.

State happiness. Participants reported the extent to which they felt *happy*, *pleased*, and *content*, which were averaged to form a composite (1 = *not at all*, 5 = *very much*; M = 3.18, SD = .93,

within-person $\omega = .89$; Barrett & Russell, 1998; Geldhof, Preacher, & Zyphur, 2014). Given that we were interested in predicting state happiness—rather than stable, dispositional happiness—happiness reports were centered within-person. Within-person-centered happiness reports were then rounded to the nearest integer, to yield a set of classification categories corresponding to each integer, which would in turn make it easier to interpret model accuracy ($M = .01$, $SD = 0.75$, range $= -3$ to 2). We collected a total of 2,272 within-person-centered happiness reports, which subsequently represented our primary criterion variable for each audio recording.

Audio recordings. Participants were asked to make two recordings each day using their smartphone. In the first recording, they described the events of their day. In the second recording, they described what was occurring in a picture taken from the Thematic Apperception Test (TAT; e.g., a child conversing with his mother, who is in bed and appears to be ill; Murray, 1943). Participants were asked to speak for approximately 60 seconds in each recording; compliance with these instructions was good on average ($M = 58.05$ seconds, $SD = 14.76$). Participants were asked to e-mail each recording to an account associated with the study.

Participants completed 3,931 unique audio recordings, for a response rate of 65%. Each audio recording was paired with a concurrent report of state happiness (i.e., two recordings per happiness report). Five audio recordings were excluded from analyses because we were not able to process them because of low audio quality (i.e., our acoustic software did not yield statistical output). This left effective sample size of 3,926 audio recordings paired with the 2,272 experience-sampling happiness reports described above (i.e., one report for every 1–2 audio recordings; $M = 7.82$ recordings per participant out of 12 possible, $SD = 3.30$, range = 1–12).

For each audio recording, we used openSMILE (Eyben, Weninger, Gross, & Schuller, 2013) to extract the 88 acoustic features included in the eGeMAPS (Eyben et al., 2016). The eGeMAPS includes summary statistics for a variety of acoustic features such as pitch, loudness, and tone (e.g., mean, *SD*, and maximum). Many of the acoustic features were highly correlated (e.g., openSMILE computes mean pitch, as well the 20th, 50th, and 80th percentile of the pitch distribution for each audio recording, with which mean pitch correlated .86 to .92). Although the 88 acoustic features in the eGeMAPS were selected based on their utility in prior work detecting emotion from acoustic cues, as well as their potential to index physiological changes in voice based on people's emotional states, it is possible that in the current audio recordings (as well as those used in Studies 2 and 3) there was insufficient variability on some of these features to allow each of them to show sufficient statistical differentiation.

Given the strong correlations observed between many acoustic features, we examined whether the 88 acoustic features could be represented by a more parsimonious set of underlying dimensions. We conducted parallel analysis and Velicer's minimum average partial, and both tests indicated that a 17-factor solution best characterized the data. We subsequently performed exploratory factor analysis with oblimin rotation, which allows factors to be correlated, because of our expectation that acoustic features would be correlated with one another (see Table 1 in online supplemental material for factor loadings). We extracted 17 acoustic factors and saved each participant's score for each factor, which subsequently represented our predictor variables for each audio recording. In this solution, the first three factors appeared to capture variability in loudness (Factor 1), relative energy or amplitude (Factor 2), and pitch (Factor 3).

Results

Analytic overview. We used three machine learning algorithms to predict within-person fluctuations in happiness from acoustic cues: (a) random forests; (b) neural networks; and (c) support vector machines. Each of these algorithms takes a set of input features (e.g., acoustic cues) and learns a mathematical function linking these input features to an outcome criterion (e.g., reports of momentary happiness). The exact mechanics of each algorithm differ somewhat. Random forests represent an ensemble of many decision trees, which subdivide the outcome criterion into its respective levels based on a series of decision rules (e.g., if pitch is above average, happiness is predicted to be 3 or greater on a Likert scale, and if loudness is also above average then happiness is predicted to be 5). Neural networks transform the values of each input feature multiple times in succession before arriving at a final prediction for the outcome; these transformations create what are called hidden layers of the neural network. Support vector machines create a series of planes in multidimensional space that maximize the distance between outcome criterion which belong to different classes (e.g., 3 vs. 4 on a Likert scale); these planes are called support vectors and are a function of a subset of the input features. We used the Scikit Learn package in Python to run each of these machine learning algorithms; more information on their implementation can be found in the Scikit Learn documentation (Scikit Learn Developers, 2018a, 2018b, 2018c). Each algorithm also has several adjustable parameters. We left these parameters set to their default values in Scikit Learn, except where noted below (see Scikit Learn documentation for default values; random forests: http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html; neural networks: http://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html#sklearn.neural_network.MLPClassifier; support vector machines: http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC).[1]

[1] Random forests differ from neural networks and support vector machines in that it is an aggregate classification method, in which a series of decision trees are fit using different random subsets of the data, and the results are averaged, reducing the noise inherent to each individual decision tree and improving overall predictive accuracy. One can, therefore, adjust the number of decision trees that are included in each random forests analysis. In theory, random forests will perform better with a greater number of trees, because the results of each individual tree will be aggregated together, thereby canceling out noise inherent to any single tree and arriving at a better overall prediction. Beyond a certain forest size, however, each additional tree has diminished marginal benefit, and very large numbers of trees can tax computer processing capacity (e.g., several of the analyses reported in this paper took more than 12 hours to complete). To determine the optimal number of trees to include in our random forests, we conducted a simulation study using our data (see pages 5–6 of online supplemental material). This simulation suggested that algorithm performance improved marginally as the number of decision trees increased to 200, before plateauing. Therefore, we set the number of trees to 200 for all subsequent random forests analyses reported in this article.

These algorithms have several advantages over conventional regression. First, they can handle many predictors at once without succumbing to model overfitting or multicollinearity. This is because each algorithm constructs a function linking the input features to the output criterion on one portion of the data (i.e., the training sample), before testing the predictive value of this function on a separate subset of the data (i.e., the testing sample) to prioritize cross-validation. Second, the algorithms can uncover complex relationships—including interactions among predictors and nonlinear links between predictors and criterion—regardless of whether they are prespecified by the researcher (i.e., a researcher does not have to enter an interaction term into a machine learning model for the algorithm to uncover that interaction).

We carried out each machine learning analysis in the following manner: For each of the 3,926 audio recordings in our data set, the corresponding within-person-centered state happiness report was the criterion variable, and the 17 acoustic factors were the predictor variables. Of the entire set of within-person-centered state happiness reports, 53% (2,075) were rounded to a score of 0 (i.e., average happiness), making this our baseline (i.e., chance) level of accuracy. This is because, if we constructed an algorithm to simply predict 0 for each recording (without using any acoustic information), it would achieve an accuracy of 53%.

For the sake of completeness, we ran two additional analyses for each machine learning algorithm, including (a) one in which we used within-person-centered acoustic factor scores as our predictors rather than raw score acoustic factors (for this model, we used the original within-person-centered state happiness reports as criteria); and (b) one in which we used raw state happiness reports as our criteria rather than within-person-centered happiness reports (for this model, we used the original raw acoustic factor scores as predictors). For this latter analysis, we rounded raw state happiness reports to the nearest integer to yield a succinct set of classification categories. A total of 42% (1,639) of raw happiness reports were at the scale midpoint of 3, making this our chance level of accuracy in this analysis.

We present results across three metrics: (a) raw accuracy, or the proportion of audio recordings for which the random forests algorithm predicted the correct happiness value; (b) Pearson correlation between the set of algorithm-predicted happiness ratings and self-report happiness ratings; and (c) Cohen's κ between the set of algorithm-predicted happiness ratings and self-report happiness ratings, which serves as an index of predictive accuracy above and beyond the base rate with which each level of happiness was reported. We also report raw accuracy for each specific integer value of self-report happiness (e.g., −1 or 0 in the analyses using within-person happiness scores; 2 or 3 in the analyses using raw happiness scores).

Except where noted, all statistics reported below for machine learning models are averages taken from 1,000 bootstrapped analyses, for which we used 10-fold cross-validation to prevent overfitting our model to one subset of our data. We report 95% confidence intervals (CIs) for all statistics, which were computed empirically based on the distribution of these bootstrapped resamples.

Can we predict state happiness from acoustic features of the voice? Our ability to predict within-person fluctuations in state happiness from acoustic features of the voice was limited across all three machine learning algorithms (see Tables 1–3). Mean predictive accuracy for random forests (56%, CI [50%, 61%]), neural networks (49%, CI [45%, 54%]), and support vector machines (53%, CI [48%, 56%]) did not meaningfully exceed chance (53%; recall that chance was set at this value because 53% of all recordings corresponded to a happiness report of 0, or average within-person centered happiness).

These results indicate that random forests were the best-performing algorithm of the three we employed. Corroborating this finding, for the random forests algorithm, the Pearson correlation between model-predicted happiness and self-report happiness (r = .13, CI [.001, .26]) was positive and statistically greater than zero, although it was weak in magnitude. Likewise, Cohen's κ between model-predicted happiness and self-report happiness (κ = .12, CI [.05, .20]) was also positive and statistically greater than zero, but again was weak in magnitude (see Table 1). This final result indicates that, although the random forests algorithm accurately predicted self-reported state happiness above what would be achieved by base rate information alone, predictive accuracy was weak. In contrast, for both neural networks and support vector machines, the Pearson correlation and Cohen's κ between model-predicted happiness and self-report happiness were lower than those values observed for random forests, and often did not statistically exceed zero (see Tables 2 and 3).

Table 1
Random Forests Analyses Predicting State Happiness From Acoustic Features (Study 1)

Model description	Accuracy	Correlation	κ
Criterion: Within-person-centered happiness Predictors: Raw acoustic features	56[a] [50, 61]	.13 [.001, .26]	.12 [.05, .20]
Criterion: Within-person-centered happiness Predictors: Within-person-centered acoustic features	53[a] [50, 56]	.05 [−.04, .17]	.02 [−.02, .05]
Criterion: Raw happiness Predictors: Raw acoustic features	48[b] [44, 51]	.21 [.13, .28]	.16 [.10, .21]

Note. N = 3,926 observations, 502 participants. Entries in each cell represent M [95% confidence interval]. Raw: Uncentered scores on a given variable. Within-person-centered: Person-mean centered scores on a given variable. Accuracy: Percentage of audio recordings for which random forests predicted the correct level of happiness. Correlation: Pearson correlation between the set of predicted happiness ratings and self-report happiness ratings. κ: Cohen's κ between the set of predicted happiness ratings and self-report happiness ratings. [a] When criterion is within-person-centered happiness, chance accuracy is 53%. [b] When criterion is raw happiness, chance accuracy is 42%.

Table 2
Support Vector Analyses Predicting State Happiness From Acoustic Features (Study 1)

Model description	Accuracy	Correlation	κ
Criterion: Within-person-centered happiness Predictors: Raw acoustic features	53[a] [48, 56]	.01 [−.10, .11]	.01 [−.003, .02]
Criterion: Within-person-centered happiness Predictors: Within-person-centered acoustic features	53[a] [46, 57]	−.06 [−.13, .04]	.001 [.00, .002]
Criterion: Raw happiness Predictors: Raw acoustic features	43[b] [41, 48]	.10 [.01, .18]	.05 [.02, .09]

Note. N = 3,926 observations, 502 participants. Entries in each cell represent M [95% confidence interval]. Raw: Uncentered scores on a given variable. Within-person-centered: Person-mean centered scores on a given variable. Accuracy: Percentage of audio recordings for which support vector machines predicted the correct level of happiness. Correlation: Pearson correlation between the set of predicted happiness ratings and self-report happiness ratings. κ: Cohen's κ between the set of predicted happiness ratings and self-report happiness ratings. [a] When criterion is within-person-centered happiness, chance accuracy is 53%. [b] When criterion is raw happiness, chance accuracy is 42%.

Examining accuracy for each individual happiness value shed additional light on each model's performance. For the random forests algorithms, although accuracy for happiness reports of 0 (i.e., average happiness) was very high (95%, CI [91%, 98%]) accuracy for happiness reports of −1 and 1 was relatively low (10 and 15%, respectively) and accuracy for −2 and 2 was 0% (see Table 3 of the online supplemental material). Neural networks showed somewhat lower accuracy for happiness reports of 0 (78%, CI [70%, 86%]) but somewhat higher accuracy for happiness reports of −1 and 1 (16 and 21%) and −2 and 2 (3 and 1%; see Table 5 of the online supplemental material). This pattern suggests that, compared to random forests, neural networks appeared to better predict values of happiness other than average happiness, but this did not result in an overall greater predictive validity when examining mean accuracy, correlation, and κ. In contrast, support vector machines showed extremely high accuracy for happiness reports of 0 (99%) and near-zero predictive accuracy for all other happiness reports (<3%; see Table 4 of the online supplemental material). This pattern suggests that, compared with random forests and neural networks, support vector machines provided little predictive validity aside from consistently predicting that people felt average levels of state happiness.

Recall that "accuracy" here is defined as when the model predicts the same value of happiness as was given via self-report. When accuracy is calculated for a given value of self-reported happiness, it partly reflects the frequency with which the algorithm predicted this value, so high accuracy could be achieved simply by an algorithm that always predicts this value (though this would lead to low accuracy for other values of self-reported happiness). Given that 0 was by far the most commonly reported happiness level and was itself reported for over half of the recordings (53%), the high accuracy for this happiness level across all three algorithms could reflect genuine insight provided by the acoustic features (i.e., most people sound about average in terms of happiness at most moments) or it could simply reflect the model frequently predicting the modal happiness level because of the absence of useful distinguishing information (i.e., using base rate information). Our data do not allow us to answer this question.

Does the random forests analysis perform better with all within-person-centered information or all raw score information? We next examined whether accuracy indices were higher in machine learning models that relied on (a) within-person-centered predictors and criteria or (b) raw score predictors and criteria (see Tables 1–3). Across each machine learning algorithm, all three accuracy indices were descriptively *lower* in a model in which within-person-centered acoustic features were used as predictors and within-person-centered happiness reports were used as criteria (e.g., for random forests models, accuracy = 53%, CI [50%, 56%];

Table 3
Neural Network Analyses Predicting State Happiness From Acoustic Features (Study 1)

Model description	Accuracy	Correlation	κ
Criterion: Within-person-centered happiness Predictors: Raw acoustic features	49[a] [45, 54]	.06 [−.03, .16]	.09 [.03, .15]
Criterion: Within-person-centered happiness Predictors: Within-person-centered acoustic features	48[a] [43, 54]	.05 [−.05, .15]	.04 [−.02, .11]
Criterion: Raw happiness Predictors: Raw acoustic features	41[b] [35, 47]	.17 [.06, .28]	.12 [.04, .20]

Note. N = 3,926 observations, 502 participants. Entries in each cell represent M [95% confidence interval]. Raw: Uncentered scores on a given variable. Within-person-centered: Person-mean centered scores on a given variable. Accuracy: Percentage of audio recordings for which neural networks predicted the correct level of happiness. Correlation: Pearson correlation between the set of predicted happiness ratings and self-report happiness ratings. κ: Cohen's κ between the set of predicted happiness ratings and self-report happiness ratings. [a] When criterion is within-person-centered happiness, chance accuracy is 53%. [b] When criterion is raw happiness, chance accuracy is 42%.

r = .05, CI [−.04, .17]; κ = .02, CI [−.02, .05]) compared with primary model (e.g., for random forests models, accuracy = 56%, CI [50%, 61%]; r = .13, CI [.001, .26]; κ = .12, CI [.05, .20]).

On the other hand, across each machine learning algorithm, all three accuracy indices were descriptively *higher* relative to chance in a model in which raw happiness reports were used as a criterion compared to our primary acoustic model (e.g., for random forests models, accuracy = 48%, CI [44%, 51%]; r = .21, CI [.13, .28]; κ = .16, CI [.10, .21]; recall that chance accuracy for this model was 42%). This suggests that our machine learning algorithms provided a small amount of predictive insight into state happiness when they used between-person (rather than purely within-person) variation in happiness and acoustic features. In the case of random forests models, the observed correlation between algorithm-predicted and self-reported happiness was r = .21, equivalent to a moderate (and typical) effect size in social psychology (Richard, Bond, & Stokes-Zoota, 2003), and significantly greater than the correlation produced by our primary random forests model. Furthermore, this correlation, as well as both the raw accuracy (48%) and κ (κ = .16) produced by this random forests model, were above-chance. Although these findings do not speak to our primary research question of whether it is possible to predict within-person fluctuations in state happiness from acoustic analysis, they do testify to the overall validity of the machine learning algorithms employed in this study: This method is capable of uncovering a link between acoustic features and happiness when it is given sufficient information (in this case, information pertaining to between-person variability).

Ruling out alternative explanations for a null effect. Given that we found limited evidence for a predictive link between acoustic features and within-person fluctuations in state happiness, we ran several additional analyses to rule out artifactual explanations for these null effects. We selected random forests for the majority of these analyses, as well as analogous tests in Studies 2 and 3, because it was consistently the best-performing machine learning algorithm in our primary analyses.

First, we examined if the null effects were because of a limitation in the type of acoustic features we extracted from the audio recordings. Although the eGeMAPS is a relatively parsimonious set of features—it consists of select acoustic features that have been empirically or theoretically linked with emotion—it is possible that a broader set of acoustic features might yield more insight into within-person fluctuations in momentary happiness. We therefore reran our primary analyses using the 2016 Interspeech Computational Paralinguistic Challenge set of 6,737 acoustic features. The Computational Paralinguistic Challenge is held annually at the Interspeech conference on spoken language processing. Importantly, prior Interspeech Challenge feature sets have been shown to slightly out-perform the eGeMAPS at detecting levels of emotional valence (analogous to levels of happiness) from audio recordings (Eyben et al., 2016). The 2016 Interspeech feature set is the most recent set available for use in openSMILE. Therefore, we reran our machine learning analyses using these acoustic features as predictors.

Second, we tested whether the limited predictive accuracy we observed was because of our decision to round within-person happiness reports to the nearest integer, which could conceivably result in the loss of valuable information. To address this issue, we reran the random forests analyses reported above while using continuous, nonrounded, within-person-centered happiness reports as our criteria (random forests can be implemented for continuous outcomes; https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html).

Third, we examined if the null effects were because of a limitation in the type of audio recordings we used. In Study 1, we combined audio recordings in which participants described the events of their day and those in which participants described a relatively neutral picture in a single analysis so as to maximize statistical power. Yet, it is possible that descriptions of neutral pictures contain little signal relevant to state happiness—particularly compared with descriptions of the events of participants' days (e.g., that may have included both happy and sad events)—and, therefore, that using them in our machine learning analysis artificially lowered accuracy. To address this issue, we reran our primary analysis separately for each type of audio recording.

Finally, we tested whether the null effects were because of a limitation in machine learning analyses. Our rationale for using machine learning was that these algorithms can find more nuanced links between predictor and outcome variables compared with regression models typically used by psychologists. Yet, in light of its poor performance, machine learning algorithms may not have been the optimal analysis for the task of predicting state happiness from acoustic features. To address this issue, we examined two alternative analytic approaches for predicting state happiness from acoustic features: (a) conventional regression and correlation and (b) asking humans.

Analyses with a larger set of acoustic features. We reran our primary machine learning analyses with all three algorithms, predicting within-person state happiness reports from raw score acoustic features taken from the 2016 Interspeech Challenge set. Unlike in our machine learning analyses using the eGeMAPS, we did not use factor analysis to reduce the dimensionality of the Interspeech feature set, because the number of features (N = 6,737) was greater than the number of cases in our data set (N = 3,926). Mean predictive accuracy for random forests (53%, CI [50%, 57%]), neural networks (40%, CI [16%, 59%]), and support vector machines (53%, CI [48%, 57%]) did not meaningfully exceed chance (53%) nor did these accuracy values exceed those found when the eGeMAPS acoustic features were used as predictors (see Tables 1–3). Pearson correlations and Cohen's κ between algorithm-predicted happiness and self-reported happiness were small in magnitude and did not exceed those found when the eGeMAPS were used as predictors (random forests: r = .03, CI [−.08, .13]; κ = .03, CI [.002, .06]; neural networks: r = −.001, CI [−.13, .13]; κ = .00, CI [−.02, .02]; support vector machines: r = .06, CI [.06, .07]; κ = .001, CI [.00, .01]).[2] These results suggest that the limited predictive validity found in our primary analyses was not because of a limitation in the relatively parsimonious eGeMAPS feature set.

Analyses with continuous outcome measure. We reran our primary analyses using the random forest regression algorithm, predicting within-person state happiness reports from raw score acoustic features. In this analysis, state happiness reports were left in their original continuous form and were not rounded to the

[2] Support vector machines analyses were run with 100 bootstrapped resamples because of limitations on computer processing capacity.

nearest integer. This analysis yielded little predictive insight (mean adjusted R^2 = .01, CI [−.02, .03]), indicating that the acoustic features explained an average of only 1% of the variance in state happiness when correcting for the large number of predictors in the model. This finding suggests that the limited predictive accuracy observed above was not due to our decision to round happiness scores to the nearest integer.

Analyses with different types of audio recordings. We reran our primary random forests model separately for audio recordings involving descriptions of the events of participants' days and audio recordings involving descriptions of neutral pictures (n = 1,963 observations for each type). Each accuracy index was higher for our original acoustic model that included both types of recordings (accuracy = 56%, CI [50%, 61%]; r = .13, CI [.001, .26]; κ = .12, CI [.05, .20]) compared with an acoustic model with only recordings of the participant retelling the day's events (accuracy = 51%, CI [47%, 55%]; r = −.03, CI [−.18, .13]; κ = .00, CI [−.05, .05]) or an acoustic model with only recordings of the participant describing a picture (accuracy = 51%, CI [42%, 56%]; r = −.01, CI [−.14, .19]; κ = .01, CI [−.04, .07]). Of course, by examining only one type of recording, we cut our sample size in half, and random forests analyses perform better with larger sample sizes because they have more information with which to learn classification distinctions. Yet, these analyses do suggest that the relative lack of predictive accuracy we observed in our primary analysis above was not because of a disproportionate lack of signal in either the daily events or picture recordings.

Can conventional analyses better predict state happiness from acoustic features? To answer this question, we performed two analyses. For each of these analyses, we did not round within-person-centered state happiness reports to the nearest integer, because general linear models are typically used with continuous outcome measures.

First, to test whether the entire set of acoustic features could collectively provide insight into state happiness, we predicted within-person-centered state happiness reports from the 17 acoustic factors that we used as predictor variables in our machine learning models, using multilevel modeling to account for the fact that state happiness reports were nested within days (i.e., participants made one state happiness report each day, which corresponded to up to two audio recordings made at the same experience-sampling assessment), and days were nested within participants. We compared a baseline model (i.e., that included only a random intercept as well as a fixed and random slope to account for any possible effect of the day on which each recording was completed) with an acoustic model (i.e., in which the 17 acoustic factors were added to the baseline model as predictors). To index model fit, we used the Akaike's Information Criterion (AIC), which penalizes more complex models and for which smaller values indicate better fit; the AIC for the acoustic model (8,133.5) was higher than for the baseline model (8,001.4), This result suggests that acoustic factors did not improve our ability to predict state happiness. To provide an estimate of effect size, we calculated approximate R^2, a statistic that reflects the proportion of reduction in the residual variance between the baseline and acoustic model (LaHuis, Hartman, Hakoyama, & Clark, 2014). Residual variance was σ^2 = .398 in the baseline model and σ^2 = .397 in the acoustic model, indicating that adding acoustic features to the model reduced residual variance by only .002%.

Next, to test whether individual acoustic features were predictive of state happiness, we computed the bivariate correlations between scores on each of the 88 acoustic features and within-person-centered state happiness reports. These correlations were very weak (average absolute value: r = .02, SD = .02; range = −.06 to .05; see Table 6 of the online supplemental material). When applying a Bonferroni correction to account for the large number of correlations we tested (α = .05/88 = .0006), only one of these 88 correlations remained significant, even though with our large sample size any correlation greater than |.054| was significant. This single significant correlation was between mean momentary happiness and mean jitter (i.e., the extent to which pitch fluctuates in consecutive speech periods; r = −.06, p = .0004). The strongest predictors of momentary happiness were mean jitter, mean mel-frequency cepstral coefficient 1 (r = .05), and seven features with absolute value correlations of r = .04 (e.g., SD of pitch, SD of jitter, and harmonics-to-noise ratio).

Together, these results suggest that conventional multilevel regression and correlational analyses involving the total set of acoustic features, as well as individual acoustic features, provide almost no insight into the link between vocal acoustics and state happiness, indicating that our failure to find substantial predictive accuracy when using machine learning algorithms was not because of a limitation inherent to this type of analysis.

Can human coders better predict state happiness from acoustic features? To test this question, we randomly selected a subset of 203 acoustic recordings drawn from 25 participants (5% of the total sample). Two research assistants were briefed on the procedure for the study, and each was then asked to predict the within-person-centered level of state happiness that corresponded to each recording in the subset (i.e., −2 to 2). We made every effort to provide coders with all of the information that a computer could conceivably use to inform a machine learning analysis. Participants were given base rate information responses across the sample (e.g., they were told that over half of the recordings corresponded to a happiness report of 0) and were encouraged to use this information when making their predictions. More important, the distribution of state happiness reports in this subsample did not significantly differ from the distribution of the overall sample (−2: 2 recordings/1%; −1: 48 recordings/24%; 0: 103 recordings/51%; 1: 50 recordings/25%; 2: 0 recordings/0%; $\chi^2(4) = 4.71$, p = .45). Coders were also instructed to listen to all recordings made by a single participant once before listening a second time to make their predictions; these instructions were meant to allow participants to better estimate state happiness around each participants' mean. Of course, the human coders also had access to information that the computer did not have, namely the content of participants' speech; in light of some work showing that the words people use are indicative of emotion (e.g., Tackman et al., 2018; cf., Sun, Schwartz, Son, Kern, & Vazire, 2018), we would predict that this imbalance should provide human coders with an advantage over the machine learning algorithm (see page 12 of online supplemental material for full coding instructions).

Each coder performed poorly at this task. Predictive accuracy was 40 and 38% for the two coders, respectively, which is well below chance in this subset of audio recordings (i.e., 51%), as well as well below the predictive accuracy achieved by our random forests and neural network models (56 and 53%, respectively). Cohen's κ between self-report happiness reports and each coder's

set of predicted happiness reports were below zero (κs = −.05 and −.08) and Pearson correlations between self-report and predicted happiness reports were near-zero (rs = −.02 and .04). Each of these results indicated worse performance than our random forests and neural network models, and analogous performance to our support vector machine models (see Tables 1–3). Furthermore, although coders were much less accurate than each machine learning algorithm at predicting happiness levels of 0 (i.e., average happiness; average coder accuracy = 66%; machine learning accuracies = 78 to .99%), accuracy at all other happiness levels did not significantly differ between coders and the machine learning algorithms.

These results together suggest that human coders were no better than our machine learning algorithms at predicting state happiness from acoustic features of the voice. If anything, human coders performed worse at this task than machine learning algorithms, primarily because they made less use of base rate information, leading to poorer accuracy at classifying average levels of state happiness. These results do not imply that the computer was substantially better than humans at predicting state happiness—after all, our machine learning analyses yielded very weak predictive accuracy on the whole—nor do they imply that human coders could never "hear happiness" in acoustic recordings under any circumstances. These results do suggest, however, that humans did not have any insight into people's relative state-level happiness based on what they said and how they said it in this sample of recordings and that our failure to find substantial predictive accuracy was, therefore, not because of a limitation in the machine learning analyses we used.

Study 2

In Study 1, we found little evidence that within-person fluctuations in state happy mood can be predicted from acoustic vocal features, using multiple machine learning algorithms, multiple acoustic feature sets, and both categorical and continuous happy mood reports as outcomes. Neither conventional regression analyses nor human coders provided more insight into state happy mood than did machine learning algorithms. Yet, in analyses that relied on between-person variance, machine learning algorithms yielded well-above-chance prediction of state happy mood from acoustic features—and effect sizes comparable to what is typically seen in social psychology—testifying to the validity of this analytic method when sufficient signal is available.

However, our ability to predict state happy mood could have been limited by the number of cases used to train our machine learning algorithms—with larger sample sizes, these algorithms have more opportunities to learn what distinguishes happier or unhappier mood. In light of this possibility, in Study 2 we harnessed an independent data set of audio recordings that was nearly 18 times as large as that which we used in Study 1.

Method

Participants. Our sample was taken from a large study in which 63,827 individuals provided experience-sampling reports of happiness and voice samples across an average of 30 days using a multiplatform smartphone application. Our final sample consisted of 19,412 participants (M_{age} = 26.96, SD = 9.34; 63% women; 91% French, 4% Belgian, 3% Swiss, 2% Other). Procedures for this study received ethical approval from the University Pompeu Fabra, Barcelona, Institutional Review Board.

Procedure. Participants volunteered for the study by downloading "58 seconds," a free mobile application for iPhone and Android phones dedicated to measuring various aspects of users' well-being through short questionnaires presented at random times throughout the day. At initial signup, participants answered demographic questions including gender, age, and country of residence. Next, participants were asked which days of the week and within what time windows they wished to receive questionnaire requests (default = 7 days/week from 9:00 a.m. to 10:00 p.m.). Participants could also customize the number of daily questionnaire requests they wanted to receive (default = 4, minimum = 1, maximum = 12). The application algorithm divided each participant's day into a number of intervals equal to the number of samples to be requested, and a random time was chosen within each interval. The minimum time between two questionnaires was set to 1 hour and random sampling was ensured through a notification system that did not require users to be connected to the Internet. New times were generated each day and were independently randomized for each participant. At each of these times, participants received a notification on their mobile phone informing them that a new questionnaire was available; some of these questionnaires included our two key tasks: reporting state happiness and making an audio recording.

State happiness. Participants were asked to rate their current happiness on a slider from 0 (*very unhappy*) to 100 (*very happy*; M = 63.4, SD = 23.8). Note that this measure differs conceptually from that used in Study 1 (and Study 3) in that it represents happiness bimodally (i.e., running from extreme unhappiness to extreme happiness) rather than unimodally (i.e., running from lack of happiness to extreme happiness). As in Study 1, however, these state happiness reports were within-person-centered, rounded to the nearest multiple of 10 and divided by 10 (e.g., a score of 53 became 5, whereas a score of 58 became 6; M = .04, SD = 1.42; range: −9 to 8). We obtained 1,541,097 state happiness reports across the entire study.

Audio recordings. Participants were asked to provide a voice sample by recording directly to the app the following sentence: *"Ceci est un échantillon de ma voix aujourd'hui"* [In English: "*This is a sample of my voice today*"]. We obtained 99,207 audio recordings across the entire study.

Because of a technical error, for each audio recording we retained accurate date stamps but not accurate time stamps. To ensure that state happiness reports and audio recordings were made at the same time, we selected all pairs of state happiness reports and audio recordings that were unique to one participant and one day. For example, if a participant made only one state happiness report and one audio recording on February 17, 2013, we could be certain that these events occurred at the same time point, and this data point would be included in our analysis. In contrast, if the same participant had made two state happiness reports and two audio recordings on that same day, we could not determine which state happiness report corresponded to which audio recording, and these data points would not be included in our analysis. This criterion left us with 71,166 pairs of corresponding state happiness reports and audio recordings.

As in Study 1, we used openSMILE to extract the 88 acoustic features included in the eGeMAPS from these 71,166 audio recordings. There were 157 (0.2%) recordings that could not be read by openSMILE, leaving 71,009 recordings for inclusion in our analysis ($M = 3.66$ per participant; $SD = 4.31$). As in Study 1, we found that a 17-factor solution best characterized these 88 features (see Table 7 in online supplemental material for factor loadings). We again saved each participant's score for each of these 17 factors to use as our primary predictor variables. As in Study 1, two of the first three factors in this solution appeared to reflect variability in loudness (Factor 1) and pitch (Factor 2), whereas Factor 3 showed its highest loading from a feature capturing the rate of speech per second in the recording.

Results

Analytic overview. We followed the same analytic strategy as in Study 1, with the following exceptions. First, of the entire set of within-person-centered state happiness reports, 39% (27,716) were rounded to a score of 0 (i.e., average happiness), making this our chance level of accuracy for analyses involving within-person-centered happiness reports. Second, 23% (16,307) of raw state happiness reports were rounded to 5, which was the most frequently reported happiness level, making this our chance level of accuracy for analyses involving raw happiness reports.

Can we predict state happiness from acoustic features of the voice? Predictive accuracy was again weak, even more so than in Study 1 (see Tables 4–6). Mean predictive accuracy did not exceed chance (39%) for random forests (38.6%, CI [37.9%, 39.5%]), neural networks (38.8%, CI [37.9%, 39.7%]), or support vector machines (39.0%, CI [38.5%, 40.0%]).[3] For the random forests algorithm, Cohen's κ between model-predicted and self-reported happiness scores did exceed zero but was extremely small in magnitude ($\kappa = .007$, CI [.002, .012]). However, Cohen's κ for neural networks and support vector machines did not statistically exceed zero (κ's = .002 and .000, respectively) and across each algorithm, the Pearson correlation between model-predicted and self-reported happiness did not exceed zero ($rs < .005$). Furthermore, although accuracy for happiness reports of 0 (i.e., average happiness) was near ceiling for each algorithm (>96%)—possibly indicating use of base-rate information—accuracy for all other happiness levels was less than 4% and frequently was 0% (see Tables 8–10 of the online supplemental material). Unlike Study 1, we did not observe large or meaningful differences in predictive accuracy across machine learning models that relied on within-person centered happiness reports and acoustic features or on raw happiness reports and acoustic features (see Table 8–10 of the online supplemental material).

Ruling out alternative explanations for a null effect. As in Study 1, we tested whether our decision to round happiness reports to the nearest integer hindered our ability to gain predictive insight into state happiness. Results again showed that random forests analyses using continuous happiness reports as a criterion did not yield meaningful predictive accuracy (see pages 20–21 of online supplemental material). Also, as in Study 1, we tested whether conventional regression and correlation provided more insight into the link between acoustic features and state happiness. Results again showed that the entire set of acoustic factors, as well as individual acoustic features, did not provide insight into momentary happiness (see pages 21–22 of online supplemental material).[4]

Study 3

In Studies 1 and 2, we found little evidence that state happy mood (measured both unimodally and bimodally) could be predicted from acoustic vocal features regardless of the type of audio used (i.e., 1-minute monologues and single-sentence utterances). These findings emerged even when we used a sample size of over 71,000 recordings to train our machine learning models (Study 2). Across both studies, conventional analyses provided no insight into state happy mood than machine learning.

However, it is possible that acoustic vocal features alone are not the best acoustic source to use for the goal of predicting state happy mood. The activities people engage in are linked to their state happy mood (e.g., Kahneman, Krueger, Schkade, Schwarz, & Stone, 2004) and many activities could shape an individual's ambient acoustic environment (e.g., socializing [vs. being alone] could increase the loudness of an audio recording). In light of this possibility, in Study 3 we tested whether we could predict state happy mood using audio recordings that involved all ambient sounds including the participants' speech, other people's speech, and sounds that naturally occurred in participants' environment.

Method

Participants. Four-hundred and thirty-four students from Washington University in St. Louis enrolled in the Personality and Interpersonal Roles Study (PAIRS; Vazire et al., 2017). Participants were compensated with money and entry into a lottery (with a higher chance of winning if they completed more ESM reports). The final sample with usable audio for this study included 283 participants (68% women; $M_{age} = 19.10$, $SD = 1.73$; 54% Caucasian, 25% Asian, 10% Black, 6% Hispanic; 5% Other). Note that portions of this data (i.e., the same participants, audio recordings, and happiness reports) have been used in other papers to test hypotheses that differ from the ones tested in the present article (e.g., Edwards & Holtzman, 2017; Sun et al., 2018; Wilson, Thompson, & Vazire, 2017; see online supplemental material for full list of papers that have used the PAIRS data set). Procedures for collection of the PAIRS data set received ethical approval from the Washington University in St. Louis Institutional Review Board.

Procedure. The current study was part of a larger project, for which a complete list of measures is available at https://osf.io/akbfj/. Participants completed an initial lab assessment during which they reported demographics and a one-item trait happiness

[3] Support vector machines analyses were run with 50 bootstrapped resamples because of limitations in computer processing capacity.

[4] Unlike Study 1, we did not conduct analyses with the Interspeech Challenge feature set because of limitations in computer processing capacity. Also, unlike Study 1, we did not conduct an analysis with human coders. However, given the extremely brief nature and homogeneous content of the audio recordings used in Study 2, as well as the near-zero predictive accuracy we observed in our primary machine learning models, we deemed it extremely unlikely that either of these analyses would yield predictive accuracy into state happy mood.

Table 4
Random Forests Analyses Predicting Happiness From Acoustic Features (Study 2)

Model description	Accuracy	Correlation	κ
Criterion: Within-person-centered happiness Predictors: Raw acoustic features	38.6[a] [37.5, 39.5]	−.001 [−.027, .026]	.007 [.002, .012]
Criterion: Within-person-centered happiness Predictors: Within-person-centered acoustic features	38.2[a] [36.9, 39.2]	.017 [−.007, .041]	.027 [.017, .036]
Criterion: Raw happiness Predictors: Raw acoustic features	22.3[b] [21.5, 23.2]	.111 [.091, .129]	.024 [.016, .031]

Note. N = 71,009 observations, 19,412 participants. Entries in each cell represent M [95% confidence interval]. Raw: Uncentered scores on a given variable. Within-person-centered: Person-mean centered scores on a given variable. Accuracy: Percentage of audio recordings for which random forests predicted the correct level of happiness. Correlation: Pearson correlation between the set of predicted happiness ratings and self-report happiness ratings. κ: Cohen's κ between the set of predicted happiness ratings and self-report happiness ratings.
[a] When criterion is within-person-centered happiness, chance accuracy is 39%. [b] When criterion is raw happiness, chance accuracy is 23%.

measure ("I am someone who is happy", 1 = *disagree strongly*, 8 = *neither agree nor disagree*, 15 = *agree strongly*; M = 11.35, SD = 2.71). For the next two weeks, a subset of participants completed ESM measures of happiness up to four times per day, while wearing an unobtrusive audio recording device (the EAR) for the first 6–8 days.

State happiness. Four times per day for 14 days, participants received a text message notification and were emailed a link to a survey that contained a one-item measure of their happiness in the target hour ("From [11 a.m.–noon/2 p.m.–3 p.m./5 p.m.–6 p.m./8 p.m.–9 p.m.], how happy were you?"; 1 = *not at all*, 3 = *neither agree nor disagree*, 5 = *very*; M = 3.45, SD = 1.00, in our final sample). As in Studies 1 and 2, these state happiness reports were within-person-centered and then rounded to the nearest integer (M = −.009, SD = 0.86, range = −3 to 2).

Audio recordings. The Electronically Activated Recorder (EAR; Mehl, 2017) was programmed to record 30 second audio snippets of participants' ambient sounds, every 9.5 minute from 7 a.m. to 2 a.m. Participants were encouraged to wear the EAR as much as possible, clipped to a waistband or the outside of their clothing (not inside a bag or pocket). Although participants had no way to tell when the EAR was recording, they were told that they could decide to leave the EAR in a different room if they did not want to wear it at any time for any reason. In addition, participants received a compact disk with their recordings so that they could listen to and erase any files that they did not want the researchers to hear. Only 99 files (0.3% of the sample) were erased, involving 15 participants.

As in Studies 1 and 2, we used openSMILE to extract the 88 acoustic features included in the eGeMAPS from all 32,757 EAR recording that took place between 11 a.m.–12:00 p.m., 2 p.m.–3 p.m., 5 p.m.–6 p.m., or 8 p.m.–9 p.m. (i.e., the hours corresponding to the state happiness reports). Eighty-seven of these recordings (0.3%) could not be read by openSMILE, leaving 32,670 for analysis. Of these, we retained 15,935 recordings for which the corresponding state happiness report was made less than 1 hour after the target hour had ended (i.e., if the state happiness report concerned 11:00 a.m.–12:00 p.m., we retained any recording for which the corresponding state happiness report was made before 1:00 p.m. on the same day). We made this choice given that people's ability to accurately recall state emotion declines in a matter of hours (Robinson & Clore, 2002). However, we reran our analyses using a more restrictive criteria—retaining recordings for which the corresponding state happiness report was made less than 15 minutes after the target hour had ended—and found similar results to the ones reported below.

As in Studies 1 and 2, we tested whether the 88 acoustic features could be represented by a more parsimonious set of dimensions and this time found that a 14-factor solution best characterized the

Table 5
Support Vector Analyses Predicting Happiness From Acoustic Features (Study 2)

Model description	Accuracy	Correlation	κ
Criterion: Within-person-centered happiness Predictors: Raw acoustic features	39.0[a] [38.5, 40.1]	.004 [−.026, .023]	.000 [−.0003, .0003]
Criterion: Within-person-centered happiness Predictors: Within-person-centered acoustic features	39.0[a] [38.1, 40.3]	.002 [−.013, .021]	.001 [.0004, .003]
Criterion: Raw happiness Predictors: Raw acoustic features	21.9[b] [21.7, 23.8]	.042 [.024, .056]	.003 [.001, .005]

Note. N = 71,009 observations, 19,412 participants. Entries in each cell represent M [95% confidence interval]. Raw: Uncentered scores on a given variable. Within-person-centered: Person-mean centered scores on a given variable. Accuracy: Percentage of audio recordings for which support vector machines predicted the correct level of happiness. Correlation: Pearson correlation between the set of predicted happiness ratings and self-report happiness ratings. κ: Cohen's κ between the set of predicted happiness ratings and self-report happiness ratings.
[a] When criterion is within-person-centered happiness, chance accuracy is 39%. [b] When criterion is raw happiness, chance accuracy is 23%.

Table 6
Neural Network Analyses Predicting Happiness From Acoustic Features (Study 2)

Model description	Accuracy	Correlation	κ
Criterion: Within-person-centered happiness Predictors: Raw acoustic features	38.8[a] [37.9, 39.7]	.004 [−.020, .028]	.002 [−.002, .006]
Criterion: Within-person-centered happiness Predictors: Within-person-centered acoustic features	37.3[a] [36.1, 38.5]	.016 [−.011, .044]	.036 [.020, .052]
Criterion: Raw happiness Predictors: Raw acoustic features	22.5[b] [21.6, 23.5]	.094 [.060, .122]	.018 [.010, .027]

Note. N = 71,009 observations, 19,412 participants. Entries in each cell represent M [95% confidence interval]. Raw: Uncentered scores on a given variable. Within-person-centered: Person-mean centered scores on a given variable. Accuracy: Percentage of audio recordings for which neural networks predicted the correct level of happiness. Correlation: Pearson correlation between the set of predicted happiness ratings and self-report happiness ratings. κ: Cohen's κ between the set of predicted happiness ratings and self-report happiness ratings. [a] When criterion is within-person-centered happiness, chance accuracy is 39%. [b] When criterion is raw happiness, chance accuracy is 23%.

data (Table 12 in online supplemental material for factor loadings). We subsequently saved each participant's score for each of these 14 factors as our primary predictor variables. As in Studies 1 and 2, the first factor in this solution appeared to capture variability in loudness, suggesting that loudness is a primary acoustic dimension on which audio recordings are classified. Factor 2 appeared to capture variability in frequency of formants 1, 2, and 3, whereas Factor 3 appeared to capture variability in relative energy or amplitude.

Next, given that each state happiness report represented a summary report of the participant's happiness during an entire hour that included multiple audio recordings, we averaged scores on each acoustic feature for all sound files within the target hour (i.e., if a participant had six audio recordings between 11:00 a.m.–12:00 p.m., we took the average of each of the 14 acoustic factor scores across those six recordings). This left us with 2,550 hourly data points (i.e., 1 hour on 1 day for 1 participant) for which we also had a state happiness report to use in our final analyses (M = 9.01 data points per participant; SD = 5.04).

Results

Analytic overview. We followed the same analytic strategy as in Studies 1 and 2, with the following exceptions. First, of the entire set of within-person-centered state happiness reports, 46% (1,164) were rounded to a score of 0 (i.e., average happiness), making this our chance level of accuracy for analyses involving within-person-centered happiness reports. Second, 35% (897) of raw state happiness reports were at the scale midpoint of 3, and this was the most frequently reported happiness level, making this our chance level of accuracy for analyses involving raw happiness reports.

Can we predict state happiness from ambient acoustic sounds? As in Studies 1 and 2, predictive accuracy was weak (see Tables 7–9). Across all three machine learning algorithms, mean accuracy did not exceed chance (46%), nor did Pearson correlation or Cohen's κ between model-predicted and self-reported happiness (random forests (accuracy = 43%, CI [37%, 53%]; r = .07, CI [−.05, .20], κ = .02, CI [−.05, .12]); neural networks: accuracy = 46%, CI [42%, 48%]; r = .04, CI [−.04, .12], κ = .002, CI [−.001, .02]), support vector machines (accuracy = 42%, CI [36%, 48%]; r = .07, CI [−.04 to .17], κ = .02, CI [−.03, .09]). Furthermore, although accuracy for happiness reports of 0 (i.e., average happiness) was very high for all three algorithms (>79%), accuracy was relatively low for happiness reports of −1 and 1 (<14% each), and less than 1% for happiness reports of −2 and 2 (see Tables 13–15 of the online supplemental material). As in Study 2, we did not observe large or meaningful differences in predictive accuracy across machine learning models

Table 7
Random Forests Analyses Predicting Happiness From Acoustic Features (Study 3)

Model description	Accuracy	Correlation	κ
Criterion: Within-person-centered happiness Predictors: Raw acoustic features	43[a] [37, 53]	.07 [−.05, .20]	.02 [−.05, .12]
Criterion: Within-person-centered happiness Predictors: Within-person-centered acoustic features	44[a] [36, 50]	.08 [−.02, .20]	.03 [−.04, .10]
Criterion: Raw happiness Predictors: Raw acoustic features	37[b] [33, 43]	.13 [.02, .29]	.06 [−.01, .14]

Note. N = 2,550 observations, 283 participants. Entries in each cell represent M [95% confidence interval]. Raw: Uncentered scores on a given variable. Within-person-centered: Person-mean centered scores on a given variable. Accuracy: Percentage of hourly data points for which random forests predicted the correct level of happiness. Correlation: Pearson correlation between the set of predicted happiness ratings and self-report happiness ratings. κ: Cohen's κ between the set of predicted happiness ratings and self-report happiness ratings. [a] When criterion is within-person-centered happiness, chance accuracy is 46%. [b] When criterion is raw happiness, chance accuracy is 35%.

Table 8
Support Vector Analyses Predicting Happiness From Acoustic Features (Study 3)

Model description	Accuracy	Correlation	κ
Criterion: Within-person-centered happiness Predictors: Raw acoustic features	42[a] [36, 48]	.07 [−.04, .17]	.02 [−.03, .09]
Criterion: Within-person-centered happiness Predictors: Within-person-centered acoustic features	46[a] [39, 52]	.03 [−.13, .10]	.01 [−.01, .02]
Criterion: Raw happiness Predictors: Raw acoustic features	37[b] [33, 43]	.14 [.05, .22]	.04 [−.02, .14]

Note. N = 2,550 observations, 283 participants. Entries in each cell represent M [95% confidence interval]. Raw: Uncentered scores on a given variable. Within-person-centered: Person-mean centered scores on a given variable. Accuracy: Percentage of hourly data points for which support vector machines predicted the correct level of happiness. Correlation: Pearson correlation between the set of predicted happiness ratings and self-report happiness ratings. κ: Cohen's κ between the set of predicted happiness ratings and self-report happiness ratings. [a] When criterion is within-person-centered happiness, chance accuracy is 46%. [b] When criterion is raw happiness, chance accuracy is 35%.

that relied on within-person centered happiness reports and acoustic features and raw score happiness reports and acoustic features (see Tables 13–15 of the online supplemental material).

Ruling out alternative explanations for a null effect. As in Study 1, we again found that random forests analyses using continuous happiness reports as a criterion did not yield meaningful predictive accuracy (see pages 20–21 of online supplemental material). Also, as in Studies 1 and 2, we again found that conventional regression and correlation analyses provided minimal insight into the link between acoustic features and state happiness (see pages 21–22 of online supplemental material). As in Study 2, we did not rerun analyses using the Interspeech Challenge feature set because of the computational burden this large feature set introduced.

In Study 3, we also examined whether the limited predictive accuracy observed above was because only a minority of audio recordings contained participants' speech. The acoustic features included in the eGeMAPS, which we used as predictors in our machine learning analysis, were developed to capture variability in emotional content of individuals' speech (Eyben et al., 2016). Therefore, it is possible that these acoustic features could not capture variability in state happiness that is potentially reflected in ambient, nonvocal sounds often included in EAR files. If so, then including these nonvocal EAR recordings could have artificially curtailed the predictive power of our machine learning analysis.

To rule out this possibility, we reran our random forests analyses using only those EAR files which contained participants' speech. Of the 15,935 audio recordings that we used in our primary analyses above, 4,447 (28%) contained speech. Following the same procedure as above, we saved each participants' scores on 14 acoustic factors for each of these recordings and averaged these factor scores within each hour. This left us with 1,606 hourly data points corresponding to a unique state happiness report to use in our final analyses (M = 5.80 data points per participant; SD = 3.64), compared with 2,550 in our primary analyses above. Of these 1,606 happiness reports, 43% (703) corresponded to a happiness report of 0 (i.e., average happiness), making this the chance level of accuracy for these analyses.

As in our primary analyses, the acoustic random forests model using only audio recordings that contained speech did not yield any predictive accuracy regarding happiness (accuracy = 40%, CI [35%, 46%]; r = .04, CI [−.10, .19]; κ = −.002, CI [−.07, .07]). It is worth noting that this analysis was hampered by a reduced sample size compared to our primary analysis which included all EAR recordings. In addition, some of these files also included speech from the people with whom the participant interacted,

Table 9
Neural Network Analyses Predicting Happiness From Acoustic Features (Study 3)

Model description	Accuracy	Correlation	κ
Criterion: Within-person-centered happiness Predictors: Raw acoustic features	46[a] [42, 48]	.04 [−.04, .12]	.002 [−.01, .02]
Criterion: Within-person-centered happiness Predictors: Within-person-centered acoustic features	42[a] [36, 46]	.09 [−.05, .21]	.03 [−.05, .11]
Criterion: Raw happiness Predictors: Raw acoustic features	35[b] [29, 41]	.11 [.01, .21]	.04 [−.04, .13]

Note. N = 2,550 observations, 283 participants. Entries in each cell represent M [95% confidence interval]. Raw: Uncentered scores on a given variable. Within-person-centered: Person-mean centered scores on a given variable. Accuracy: Percentage of hourly data points for which neural networks predicted the correct level of happiness. Correlation: Pearson correlation between the set of predicted happiness ratings and self-report happiness ratings. κ: Cohen's κ between the set of predicted happiness ratings and self-report happiness ratings. [a] When criterion is within-person-centered happiness, chance accuracy is 46%. [b] When criterion is raw happiness, chance accuracy is 35%.

which could have been less diagnostic of participants' happiness than participants' own speech. Nevertheless, the fact that we saw no improvement in predictive accuracy in these auxiliary analyses—along with the null findings in Studies 1 and 2 involving direct speech—suggests that the lack of vocal acoustic information in the EAR recordings was likely not the primary factor limiting predictive accuracy in the primary analyses.

General Discussion

Across three studies, we found little evidence that fluctuations in state happy mood could be predicted from acoustic analysis. These results emerged despite using happy mood self-reports obtained via intensive smartphone-based experience-sampling as our criteria, multiple sets of acoustic features as our predictors (one of which was tailor-made to detect emotion), and three cutting-edge machine learning algorithms as our analytic tools. This result also emerged regardless of whether happy mood scores were treated as categorical or continuous outcome measures and regardless of whether we used as predictors acoustic features capturing the voice (Studies 1 and 2) or ambient sounds (Study 3). Furthermore, in Study 1, audio recordings were made in private—and, therefore, participants had no externally imposed reasons to enact a certain emotional tone—whereas the recordings in Studies 2–3 may have taken place in public settings where display rules or other social norms could have shaped emotional expression in the voice. Yet, regardless of these differences in the social context in which recordings were made, predictive accuracy remained near-zero. We also ruled out alternative explanations for these null effects, showing that conventional regression and correlation analyses did not yield any predictive insight into state happy mood (Studies 1–3), that human coders could not accurately predict state happy mood (Study 1), and that machine learning analysis yielded above-chance accuracy when relying on between-person variability in happy mood. We observed these null effects despite collecting large samples, including one that was orders of magnitude greater than is typical for psychological studies.

What Did We Learn?

The present research suggests that, despite popular claims that computers will soon be able to hear our emotions, it does not yet seem possible to automatically predict fluctuations in one of the most basic state emotional experiences—happy mood—through acoustic analysis. At first blush, this conclusion appears to stand in contrast to two more encouraging lines of work in this broad conceptual space. First, recent psychological research has made groundbreaking progress in harnessing emergent technologies and big data methodologies other than acoustic analysis to automatically predict *community* and *trait* happiness using social media data (e.g., Dodds et al., 2011; Schwartz et al., 2013, 2014). Second, recent work has shown that distinct emotions have distinct vocal signatures, including nonverbal vocal bursts (e.g., sighs, gasps, and grunts; Cordaro et al., 2016; Sauter et al., 2010) and distinct profiles of acoustic features conveyed via speech (e.g., pitch, energy; Banse & Scherer, 1996; Wallbott & Scherer, 1986). Furthermore, machine learning techniques can help automatically distinguish between distinct emotion states when they are portrayed by actors via speech (Laukka et al., 2014) and manifested as extreme, clinical varieties in conversations (Karam et al., 2014).

Upon closer examination, however, the present findings are reconcilable with these two lines of prior work. Our findings suggest that it is far more challenging to predict moment-to-moment variability in levels of happy mood, compared with predicting stable community or trait happiness, or compared to differentiating between distinct emotions. This is likely because the acoustic differences between the same individual experiencing high versus low happy mood may be very subtle (i.e., within-person variability), whereas differences aggregated across time between two individuals who are typically high versus low in happiness (i.e., between-person variability), and differences between distinct emotions such as anger and happiness, may be much larger. Consistent with this perspective, recent work has shown that linguistic analysis of speech content during everyday life (as opposed to acoustic properties of speech, as in the current research) provides a surprisingly small amount of predictive insight into within-person fluctuations in state happiness (Sun et al., 2018). In contrast, the predictive correlation obtained in one prior study of *trait* happiness (e.g., $r = .17$; Kosinski et al., 2013) was similar to that obtained in Study 1 when raw score acoustic features were used to predict raw score happy mood reports (i.e., an analysis involving between-person information; $r = .21$; see Table 1).

It is also worth noting that prior work has shown that fluctuations in valence (analogous to fluctuations in happy mood) may covary relatively weakly with specific acoustic properties, compared to other emotion dimensions (e.g., arousal, potency; Laukka, Juslin, & Bresin, 2005; Pereira, 2000). This may be because variability in valence—particularly pleasant levels of valence that were the focus of the current study—may primarily reflect the experience of one of multiple distinct positive emotions, each of which may be associated with distinct nonverbal vocal profiles (e.g., amusement, relief, and contentment; Sauter & Scott, 2007). The inability of current algorithms to distinguish between levels of happy mood, therefore, does not preclude future machine learning research from distinguishing between distinct emotions such as anger, fear, or disgust, which may be better differentiated via acoustic cues.

In summary, although it does not seem possible to predict fluctuations in state happy mood via acoustic analysis at the present time, our predictive goal diverged from prior work in this space in several ways that likely made it less tractable.

Constraints on Generality and Future Directions

The samples used in Studies 1–3 consisted primarily of undergraduate students and young adults, albeit from three distinct cultural contexts (Canada, France, and Missouri). We have no reason to believe that the acoustic vocal features associated with fluctuations in happy mood would manifest differently in older adults, or individuals from different cultures and, therefore, expect that our findings would replicate across different participant samples. We also replicated our findings using two different strategies for assessing happy mood (momentary reports in Studies 1–2; retrospective reports in Study 3), two different acoustic feature sets (one relatively parsimonious and one extremely large) and three different machine learning algorithms (each of which has been viewed as cutting-edge in the present day or recent past). We therefore expect that our findings would replicate across alterna-

tive means of assessing state happy mood, alternative acoustic feature sets, and alternative machine learning algorithms. However, it is entirely possible that features other than, or in combination with, acoustics may provide more insight into fluctuations in state happy mood (e.g., speech content; cf. Kross et al., in press).

The primary constraint on the generality of these findings—and one that points to important future work—concerns the specific properties of audio recordings we used. In Studies 1–3, we collected recordings that (a) often involved mundane vocalizations that likely did not reflect intense emotion and (b) were very brief in duration (i.e., thin slices). In fact, given that the vast majority of happy mood reports across all three studies fell at or within one point of participants' within-person averages, one could argue that the audio recordings used in the present research captured relatively subtle fluctuations in happy mood. This implies that there may have been insufficient emotional signal in the acoustic recordings we used to support predictive accuracy obtained via machine learning algorithms.

An important future direction is to conduct similar predictive analyses via machine learning while using as input audio recordings that convey more intense happy or unhappy mood (as in Karam et al., 2014) or convey distinct emotions altogether (as in Laukka et al., 2014). Of course, because intense emotional experiences are relatively rare (Diener, Kanazawa, Suh, & Oishi, 2015), audio recordings with more extreme happy mood or distinct emotional reactions may be less representative of the emotional states that people experience in their everyday lives. In fact, it may be that most moments in everyday life are likely somewhat mundane and that predicting fluctuations in happy mood may be quite difficult even with better technology. Even so, if the goal is to automatically assess fluctuations in state happy mood at scale—as was the goal of the present work—it will be important to demonstrate this predictive power on a set of audio recordings that is representative of the intensity and range of fluctuations in happy mood typically experienced in daily life.

Conclusion

The present null findings do not preclude the possibility that automatically predicting fluctuations in emotional states such as happy mood will someday be possible, particularly considering other data sources that could be used to accomplish this task. The present null findings do suggest that affective scientists are not ready to automatize the detection of "happy moments." Yet, we look forward to future work that will challenge this conclusion and continue to make progress in acoustic analysis of emotion and automated emotion detection.

References

Affectiva. (2018). *Affectiva automotive AI*. Retrieved from https://www.affectiva.com/product/affectiva-automotive-ai/

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders: DSM–5* (5th ed.). Arlington, VA: Author.

Baird, B. M., & Lucas, R. E. (2011). ". . . And how about now?": Effects of item redundancy on contextualized self-reports of personality. *Journal of Personality, 79,* 1081–1112. http://dx.doi.org/10.1111/j.1467-6494.2011.00716.x

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology, 70,* 614–636. http://dx.doi.org/10.1037/0022-3514.70.3.614

Beck, E. D., & Jackson, J. J. (2018). *Consistency and change in idiographic personality: A longitudinal ESM network study*. Manuscript submitted for publication. Retrieved from https://psyarxiv.com/pb92q/

Beyond Verbal. (2018). *Virtual private assistant*. Retrieved from http://www.beyondverbal.com/vpa/

Breil, S. M., Geukes, K., Wilson, R. E., Nestler, S., Vazire, S., & Back, M. (2018). *Zooming into real-life extraversion–How personality and situation shape sociability in social interactions*. Manuscript submitted for publication.

Colman, D. E., Vineyard, J., & Letzring, T. D. (2018). Exploring beyond simple demographic variables: Differences between traditional laboratory samples and crowdsourced online samples on the Big Five personality traits. *Personality and Individual Differences, 133,* 41–46. http://dx.doi.org/10.1016/j.paid.2017.06.023

Conner, T. S., Tennen, H., Fleeson, W., & Barrett, L. F. (2009). Experience sampling methods: A modern idiographic approach to personality research. *Social and Personality Psychology Compass, 3,* 292–313. http://dx.doi.org/10.1111/j.1751-9004.2009.00170.x

Cordaro, D. T., Keltner, D., Tshering, S., Wangchuk, D., & Flynn, L. M. (2016). The voice conveys emotion in ten globalized cultures and one remote village in Bhutan. *Emotion, 16,* 117–128. http://dx.doi.org/10.1037/emo0000100

Diener, E., Kanazawa, S., Suh, E. M., & Oishi, S. (2015). Why people are in a generally good mood. *Personality and Social Psychology Review, 19,* 235–256. http://dx.doi.org/10.1177/1088868314544467

Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., & Danforth, C. M. (2011). Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PLoS ONE, 6,* e26752. http://dx.doi.org/10.1371/journal.pone.0026752

Edwards, T., & Holtzman, N. S. (2017). A meta-analysis of correlations between depression and first person singular pronoun use. *Journal of Research in Personality, 68,* 63–68. http://dx.doi.org/10.1016/j.jrp.2017.02.005

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., Andre, E., Busso, C., . . . Truong, K. P. (2016). The Geneva Minimalist Acoustic Parameter Set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing, 7,* 190–202. http://dx.doi.org/10.1109/TAFFC.2015.2457417

Eyben, F., Weninger, F., Gross, F., & Schuller, B. (2013). Recent developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor. *MM '13 Proceedings of the 21st ACM international conference on Multimedia* (pp. 835–838). New York, NY: ACM. http://dx.doi.org/10.1145/2502081.2502224

Feldman Barrett, L., & Russell, J. A. (1998). Independence and bipolarity in the structure of current affect. *Journal of Personality and Social Psychology, 74,* 967–984. http://dx.doi.org/10.1037/0022-3514.74.4.967

Finnigan, K. M., & Vazire, S. (2018). The incremental validity of average state self-reports over global self-reports of personality. *Journal of Personality and Social Psychology, 115,* 321–337. http://dx.doi.org/10.1037/pspp0000136

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19,* 72–91. http://dx.doi.org/10.1037/a0032138

Harari, G. M., Lane, N. D., Wang, R., Crosier, B. S., Campbell, A. T., & Gosling, S. D. (2016). Using smartphones to collect behavioral data in psychological science: Opportunities, practical considerations, and challenges. *Perspectives on Psychological Science, 11,* 838–854. http://dx.doi.org/10.1177/1745691616650285

Joel, S., Eastwick, P. W., & Finkel, E. J. (2017). Is romantic desire predictable? Machine learning applied to initial romantic attraction. *Psychological Science, 28,* 1478–1489. http://dx.doi.org/10.1177/0956797617714580

Kahneman, D., Krueger, A. B., Schkade, D. A., Schwarz, N., & Stone, A. A. (2004). A survey method for characterizing daily life experience:

The day reconstruction method. *Science, 306,* 1776–1780. http://dx.doi.org/10.1126/science.1103572

Karam, Z. N., Provost, E. M., Singh, S., Montgomery, J., Archer, C., Harrington, G., & Mcinnis, M. G. (2014, May). *Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech*. In International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy.

Kassam, K. S., & Mendes, W. B. (2013). The effects of measuring emotion: Physiological reactions to emotional situations depend on whether someone is asking. *PLoS ONE, 8,* e64959. http://dx.doi.org/10.1371/journal.pone.0064959

Knight, W. (2016, June). Amazon working on making Alexa recognize your emotions. *MIT Technology Review*. Retrieved from https://www.technologyreview.com/s/601654/amazon-working-on-making-alexa-recognize-your-emotions/

Kosinski, M., Stillwell, D., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of the United States of America, 110,* 5802–5805. http://dx.doi.org/10.1073/pnas.1218772110

Kross, E., Verduyn, P., Boyer, M., Drake, B., Gainsburg, I., Vickers, B., . . . Jonides, J. (2018). Does counting emotion words on online social networks provide a window into people's subjective experience of emotion? A case study on Facebook. *Emotion*. Advance online publication. http://dx.doi.org/10.1037/emo0000416

Kuppens, P., Oravecz, Z., & Tuerlinckx, F. (2010). Feelings change: Accounting for individual differences in the temporal dynamics of affect. *Journal of Personality and Social Psychology, 99,* 1042–1060. http://dx.doi.org/10.1037/a0020962

LaHuis, D. M., Hartman, M. J., Hakoyama, S., & Clark, P. C. (2014). Explained variance measures for multilevel models. *Organizational Research Methods, 17,* 433–451. http://dx.doi.org/10.1177/1094428114541701

Laukka, P., Juslin, P., & Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognition and Emotion, 19,* 633–653. http://dx.doi.org/10.1080/02699930441000445

Laukka, P., Neiberg, D., & Elfenbein, H. A. (2014). Evidence for cultural dialects in vocal emotion expression: Acoustic classification within and across five nations. *Emotion, 14,* 445–449. http://dx.doi.org/10.1037/a0036048

Lieberman, M. D., Inagaki, T. K., Tabibnia, G., & Crockett, M. J. (2011). Subjective responses to emotional stimuli during labeling, reappraisal, and distraction. *Emotion, 11,* 468–480. http://dx.doi.org/10.1037/a0023503

Lyubomirsky, S., & Lepper, H. S. (1999). A measure of subjective happiness: Preliminary reliability and construct validation. *Social Indicators Research, 46,* 137–155. http://dx.doi.org/10.1023/A:1006824100041

Mehl, M. R. (2017). The electronically activated recorder (EAR): A method for the naturalistic observation of daily social behavior. *Current Directions in Psychological Science, 26,* 184–190. http://dx.doi.org/10.1177/0963721416680611

Mehl, M. R., & Conner, T. S. (Eds.). (2013). *Handbook of research methods for studying daily life*. New York, NY: Guilford Press.

Murray, H. A. (1943). *Thematic apperception test*. Cambridge, MA: Harvard University Press.

Nelson, B. W., & Allen, N. B. (2018). Extending the passive-sensing toolbox: Using smart-home technology in psychological science. *Perspectives on Psychological Science, 13,* 718–733. http://dx.doi.org/10.1177/1745691618776008

Pereira, C. (2000). Dimensions of emotional meaning in speech. *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion* (pp. 25–28). Retrieved from https://www.isca-speech.org/archive_open/speech_emotion/spem_025.html

Pew Research Center. (2018, February 5). *Mobile fact sheet*. Retrieved from http://www.pewinternet.org/fact-sheet/mobile/

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7,* 331–363. http://dx.doi.org/10.1037/1089-2680.7.4.331

Robinson, M. D., & Clore, G. L. (2002). Episodic and semantic knowledge in emotional self-report: Evidence for two judgment processes. *Journal of Personality and Social Psychology, 83,* 198–215. http://dx.doi.org/10.1037/0022-3514.83.1.198

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called *emotion*: Dissecting the elephant. *Journal of Personality and Social Psychology, 76,* 805–819. http://dx.doi.org/10.1037/0022-3514.76.5.805

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences of the United States of America, 107,* 2408–2412. http://dx.doi.org/10.1073/pnas.0908239106

Sauter, D. A., & Scott, S. K. (2007). More than one kind of happiness: Can we recognize vocal expressions of different positive states? *Motivation and Emotion, 31,* 192–199. http://dx.doi.org/10.1007/s11031-007-9065-x

Scherer, K. R. (1986). Vocal affect expression: A review and a model for future research. *Psychological Bulletin, 99,* 143–165. http://dx.doi.org/10.1037/0033-2909.99.2.143

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., . . . Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE, 8,* e73791. http://dx.doi.org/10.1371/journal.pone.0073791

Schwartz, H. A., Eichstaedt, J., Kern, M. L., Park, G., Sap, M., Stillwell, D., . . . Ungar, L. (2014). Towards assessing changes in degree of depression through Facebook. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (pp. 118–125). Baltimore, MD: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/W/W14/W14-3214

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54,* 93–105. http://dx.doi.org/10.1037/0003-066X.54.2.93

Scikit Learn Developers. (2018a). *Neural network models*. Retrieved from http://scikit-learn.org/stable/modules/neural_networks_supervised.html

Scikit Learn Developers. (2018b). *Random forest classifier*. Retrieved from http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

Scikit Learn Developers. (2018c). *Support vector machines*. Retrieved from http://scikit-learn.org/stable/modules/svm.html

Shrout, P. E., Stadler, G., Lane, S. P., McClure, M. J., Jackson, G. L., Clavél, F. D., . . . Bolger, N. (2018). Initial elevation bias in subjective reports. *Proceedings of the National Academy of Sciences of the United States of America, 115,* E15–E23. http://dx.doi.org/10.1073/pnas.1712277115

Solomon, B. C., & Vazire, S. (2016). Knowledge of identity and reputation: Do people have knowledge of others' perceptions? *Journal of Personality and Social Psychology, 111,* 341–366. http://dx.doi.org/10.1037/pspi0000061

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods, 14,* 323–348. http://dx.doi.org/10.1037/a0016973

Sun, J., Schwartz, H. A., Son, Y., Kern, M. L., & Vazire, S. (2018). *The language of well-being: Tracking within-person emotion fluctuations through everyday speech*. Manuscript submitted for publication.

Sun, J., & Vazire, S. (in press). Do people know what they're like in the moment? *Psychological Science*. Retrieved from https://psyarxiv.com/sg5aw/

Tackman, A. M., Sbarra, D. A., Carey, A. L., Donnellan, M. B., Horn, A. B., Holtzman, N. S., . . . Mehl, M. R. (2018). Depression, negative emotionality, and self-referential language: A multi-lab, multi-measure,

and multi-language-task research synthesis. *Journal of Personality and Social Psychology*. Advance online publication. http://dx.doi.org/10.1037/pspp0000187

Vazire, S., Wilson, R. E., Solomon, B. C., Bollich, K. L., Harris, K., Weston, S. J., . . . Jackson, J. J. (2017). Personality and Interpersonal Roles (PAIRS). *Study in progress*. Retrieved from https://osf.io/akbfj/

Vogels, W. (2018, January). Voice is about to fix our love-hate relationship with machines. *WIRED Magazine*. Retrieved from https://www.wired.co.uk/article/google-home-amazon-alexa-apple-homepod-voice-recognition

Wallbott, H. G., & Scherer, K. R. (1986). How universal and specific is emotional experience? Evidence from 27 countries on five continents. *Social Sciences Information, 25,* 763–795. http://dx.doi.org/10.1177/053901886025004001

Wang, Y., & Kosinski, M. (2018). Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology, 114,* 246–257. http://dx.doi.org/10.1037/pspa0000098

Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological Bulletin, 98,* 219–235. http://dx.doi.org/10.1037/0033-2909.98.2.219

Wilson, R. E., Harris, K., & Vazire, S. (2015). Personality and friendship satisfaction in daily life: Do everyday social interactions account for individual differences in friendship satisfaction? *European Journal of Personality, 29,* 173–186. http://dx.doi.org/10.1002/per.1996

Wilson, R. E., Thompson, R. J., & Vazire, S. (2017). Are fluctuations in personality states more than fluctuations in affect? *Journal of Research in Personality, 69,* 110–123. http://dx.doi.org/10.1016/j.jrp.2016.06.006

Received July 11, 2018
Revision received December 10, 2018
Accepted December 12, 2018 ■