Historical heterogeneity predicts smiling: Evidence from large-scale observational analyses

Jeffrey M. Girard[1] and Daniel McDuff[2]
[1] University of Pittsburgh, Pittsburgh, PA 15260
[2] Microsoft Research, Redmond, WA 98052

***Abstract*— Facial behavior is a valuable source of information about an individual's feelings and intentions. However, many factors combine to influence and moderate facial behavior including personality, gender, context, and culture. Due to the high cost of traditional observational methods, the relationship between culture and facial behavior is not well-understood. In the current study, we explored the sociocultural factors that influence facial behavior using large-scale observational analyses. We developed and implemented an algorithm to automatically analyze the smiling of 866,726 participants across 31 different countries. We found that participants smiled more when from a country that is higher in individualism, has a lower population density, and has a long history of immigration diversity (i.e., historical heterogeneity). Our findings provide the first evidence that historical heterogeneity predicts actual smiling behavior. Furthermore, they converge with previous findings using self-report methods. Taken together, these findings support the theory that historical heterogeneity explains, and may even contribute to the development of, permissive cultural display rules that encourage the open expression of emotion.**

I. Introduction

Facial behavior is a highly visible avenue of affective and interpersonal communication that can signal information about feelings, cognitive states, and action tendencies [1]. As a result, the face provides a powerful behavioral-lens through which to view numerous scientific topics, from basic research on emotion and personality to applied research on assessing and treating psychopathology [2]. There is also considerable interest in facial behavior within the public, health-care, defense/security, and consumer-goods sectors, where various endeavours would be facilitated by a deeper understanding of individuals' experiences and intentions.

However, the face does not provide a perfect read-out of an individual's inner world [3]; rather, many factors combine to influence facial behavior. In addition to the emotions, cognitions, and intentions that are often of primary interest, an individual's behavior may be influenced by his or her personality, mental health, age, gender, and ethnicity. For example, introverts tend to smile less than extraverts [4], men tend to smile less than women [5], and currently-depressed patients tend to smile less than recovered patients [6].

Behavior is also contextualized within a specific situation that has its own social and affective characteristics. These characteristics may, in turn, promote or inhibit different behaviors. For example, the distribution and interpretation of behavior is likely to differ between celebratory and mournful situations, as well as between cooperative and competitive ones. The mere presence of other people also may be impactful, as people tend to smile more in social situations than in solitary ones, even when reporting similar levels of positive affect (e.g., happiness or amusement) [7]–[9].

On a larger scale, behavior is also embedded within the context of *culture*, which is a shared system of meaning and information that maintains social order by providing values and norms for behavior, thought, and emotion [10], [11]. These values and norms can have a profound impact on how facial behavior is displayed and interpreted. For example, some cultures value the free expression of emotion, while others value its careful curation [11], [12]. The probability, and therefore the meaning, of displaying an emotion-related behavior in two such cultures would be quite different.

Currently, the relationship between culture and facial behavior is not well-understood. Most of the research on this topic has been limited by either relying on self-reported facial behavior, which may not represent actual behavior, or by observing a small number of participants from a small number of countries, which may yield results that fail to generalize. Further research is needed to quantify the extent to which facial behavior differs across cultures, as well as to investigate which cultural factors explain such differences.

A. Cross-cultural Differences in Facial Behavior

Klineberg [13] proposed several ways in which culture might influence expressive behavior. First, culture might dictate the meaning/interpretation of a behavior. Second, culture might influence what emotions, and therefore what expressions, occur in a given situation. Finally, culture might determine whether expressions are permitted, inhibited, or exaggerated in a given situation. The term "display rule" was later coined for this form of normative control [14].

Display rules have been studied using a variety of methods ranging from the rating of posed images (e.g., how appropriate would it be to make this expression?) to self-reported responses to hypothetical situations. In the largest study of display rules to date, Matsumoto et al. [11] surveyed 5,361 participants from 32 countries about what they "should do" if they felt different emotions in different situations. In this study, it was found that participants from countries higher in individualism endorsed greater expression of emotions (especially positive emotions). Individualism captures the degree of independence versus interdependence a culture

Both authors contributed equally to this study. The efforts of J. M. Girard were supported in part by a grant from the NSF (IIS-1418026).

978-1-5090-4023-0/17/$31.00

maintains among its members [15], and it has been argued that the free expression of emotions has greater importance in cultures that are higher in individualism [12], [16].

This same data was recently reanalyzed by Rychlowska et al. [17], who used multiple regression to predict each country's average endorsement of emotional expression using a larger set of sociocultural variables. In addition to individualism, it was found that the heterogeneity of a country's historical immigration explained unique variance in expressivity norms. That is, individuals in more historically diverse cultures (which had received immigrants from more countries over the past 500 years) believed that emotions should be openly expressed rather than dissimulated.

This finding supports the theory that permissive display rules and amplified expressivity increase the accuracy of communication and the building of trust, both of which are more important in historically heterogeneous societies (in which disparate cultures have routinely collided) than in historically homogeneous societies (in which common practices and rules guide expectations) [17]–[19].

While self-reports can reveal how emotions and facial behaviors are perceived by a culture, their results do not necessarily reflect actual behavior. A direct test of cross-cultural differences in facial behavior would require observational measures of participants from different cultures. Unfortunately, observational research is very expensive and time-consuming—a problem that is only compounded when a study is scaled across multiple countries. Thus, previous observational studies have been restricted to a small number of participants and/or a small number of countries.

Friesen [20] compared the facial behavior of 25 Japanese and 25 American university students while watching stressful films; the results have been widely interpreted as evidence of a Japanese display rule that negative emotions should be masked in the presence of authority figures [21]. Waxer [22] similarly compared the nonverbal behavior of 30 contestants on American TV game shows and 30 contestants on Canadian TV game shows; the results were interpreted as evidence that American display rules encourage greater expression of emotion than Canadian display rules. Later studies compared the facial behavior of children (i.e., infants and preschoolers) from several Western and Asian countries; their results suggest that cultural differences in facial behavior begin quite early in life and are related to both temperamental differences and family environmental factors [23]–[26].

These country-comparison studies demonstrated that facial behavior does indeed vary across cultures, but their inclusion of a small number of countries makes it difficult to know which cultural factors explain the cross-cultural differences.

Matsumoto, Willingham, & Olide [27] were the first to examine the relationship between observed facial behavior and sociocultural variables. By examining the facial behavior of 84 Olympic athletes from all over the world [28], they were able to identify predictors of greater facial expressivity. Specifically, participants were more expressive if they came from countries with higher affluence, population density, and individualism. With participants spanning 35 different countries, this study included far more cultural diversity than any previous cross-cultural observational study. However, with an average of only 2.4 participants per country, this impressive breadth of coverage came at the cost of depth.

Recently, McDuff, Girard, & el Kaliouby [29] used techniques from computer science to conduct the first truly large-scale observational study of facial behavior. The behavior of 740,984 participants from 12 different countries was analyzed using computer algorithms. These algorithms could detect, with high accuracy, the extent to which participants made smiling and brow-furrowing expressions while watching television ads. It was found that participants' facial behavior was influenced by a combination of factors including culture (i.e., individualism), gender, and context.

In terms of cultural factors, McDuff et al. [29] found that the influence of individualism on brow-furrowing was straight-forward: participants from countries higher in individualism displayed more brow-furrowing overall. In contrast, the influence of individualism on smiling depended on context: participants from countries higher in individualism smiled more in the context of a market research facility, whereas participants from countries lower in individualism smiled more in their own homes. One limitation of this study was that the inclusion of only 12 different countries restricted the number of cultural factors that could be examined. Thus, no study has yet tested the influence of historical immigration heterogeneity on observed facial behavior or replicated Matsumoto et al.'s [27] findings that higher affluence and population density predict greater expressivity.

B. The Current Study

We build upon the work in McDuff et al. [29] and apply the same methodology to examine a wider range of cultural factors in a larger and more diverse sample. To maximize comparability across different countries and to streamline our hypotheses, we limit our sample to participants who were observed in the context of a market research facility and focus our efforts on predicting each country's average amount of observed smiling. The smile is a particularly important and prevalent facial behavior that can communicate information about emotion, affiliation, and dominance [30], [31]. For cultural factors, we include measures of individualism, affluence, population density, urbanization, historical heterogeneity, and present-day ethnic diversity.

To summarize our methodology, first, we used an Internet-based framework to collect videos of participants while they watched televisions ads in a market research facility. This setting enabled us to collect a large amount of data in many different countries at low cost, as corporations were willing to compensate participants for viewing their ads. Then, we developed and implemented a facial coding algorithm to automatically measure the extent to which participants were smiling in each video. Finally, we averaged the amount of observed smiling in each country and tried to predict it using a number of theoretically relevant cultural factors.

On the basis of theory and previous research, we hypothesize that (1) different countries will have very different levels

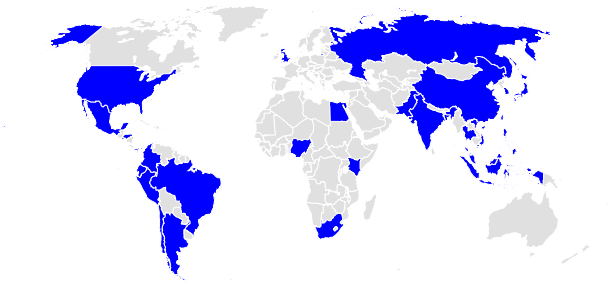

Fig. 1. World map with the 31 included countries shaded in dark blue.

of observed smiling, (2) countries with higher individualism will have more observed smiling, (3) countries with higher historical immigration heterogeneity will have more observed smiling, (4) countries with higher affluence will have more observed smiling, and (5) countries with higher population density will have more observed smiling.

II. Methods

A. Data Collection

A scalable Internet-based framework [32] was used to collect videos of participants' facial responses to ad content. Videos were recorded via webcam at a framerate of 14 frames per second and at a resolution of 320×240 pixels. Participants were recruited through market research panels and took part using computers in market research facilities. At the start of each session, participants provided consent to participate in the study and to have their facial behavior recorded. They were monetarily compensated at a local rate similar to that for a typical 30 min market research survey.

Over a period of 5 years, we collected videos of 1,494,079 participants in 56 different countries. In the current study, we focus on countries that had 30 or more different ad stimuli and participants that had facial tracking in more than 50 % of video frames; this yielded a final count of 866,726 participants from 31 countries (see Table I and Figure 1). Of the participants who provided gender information, 73 % were female; gender ratios were similar across all included countries ($SD = 16\,\%$). Videos from 12 of the 31 countries were also included in McDuff et al.'s [29] 'facility' subsample.

During each session, participants watched between one and ten video ads. These ad stimuli were selected from typical television content in each country and represented a wide cross-section of major brands and products. Using country-specific stimuli conferred several advantages over the presentation of identical stimuli to all participants, including that it allowed us to circumvent language barriers and to avoid confusion caused by culture-specific references.

A total of 7,875 different ad stimuli were used (Table I). In all countries, ads from the following categories were selected: beverages, groceries, personal care, home goods, services, and retail. The average duration of the ads, weighted by the number of participants who viewed each ad, was 32.8 s ($SD = 16.0$ s). With such a large number of stimuli, the influence of any single ad within a country was minimized.

B. Automated Smile Coding

Supervised learning was used to create an automated smile detection algorithm. A large set of webcam videos was manually coded for this purpose. Coding was provided by a group of 20 coders who were trained using material from the Facial Action Coding System (FACS) manual [33]. A minimum of three coders labeled each video frame for the presence or absence of AU 12 (i.e., the lip corner puller). A fourth coder, who had been certified by passing the official FACS Final Test, checked the agreement and quality of the labels before they were approved. If a label failed this quality check, the video was relabeled by the original coders.

The reliability of the human coding was measured in terms of video-level *base rates* (i.e., the proportion of time that a participant smiled in each video) and frame-level *occurrence* (i.e., the presence or absence of a smile in each frame). The video-level reliability of the manual codes, as measured by an intraclass correlation coefficient [34], was ICC(A,1) = .91; the frame-level reliability, as measured by the free-marginal kappa coefficient [35] or S score [36], was $S = .81$.

The manually coded videos were partitioned into fully independent *training*, *validation*, and *testing* sets. The training set was a random sample of 80,000 labeled video frames from 4,000 unique individuals, and the validation set was a random sample of 10,000 frames from 2,500 unique individuals. The classes were balanced in the training and validation sets such that half of the frames were smiles and the other half were non-smiles. The sampling was designed to maximize number of unique individuals in both the smile and non-smile subsets. The testing set was a random sample of 900,000 video frames from 1,333 unique individuals; the classes were left unbalanced in this set to better reflect the domain of application. To our knowledge, this is the largest and most diverse FACS-coded dataset in existence.

Using the training set, we trained a two-class support vector machine (SVM) with a Nyström-approximated radial basis function (RBF) kernel. Kernelized SVMs are effective at building discriminative models for complicated non-linear classification tasks, but incur high computational costs as training samples are added. To circumvent this cost, we used the Nyström method to find an approximate embedding by selecting a random subset of the training samples [37].

Using the validation data, we optimized the following parameters: the number of samples used in the Nyström approximation ($N_s \in \{200, 500, 1000, 2000\}$), the SVM cost parameter ($C \in \{0.01, 0.1, \ldots, 100\}$), and the RBF spread parameter ($\gamma \in \{0.01, 0.1, \ldots, 100\}$). Further details about the implementation and validation of this approach are provided in previous publications [29], [37].

The performance of our smile detection algorithm (i.e., its agreement with human coding) was evaluated in two ways. First, we applied the trained and optimized algorithm to the reserved testing set. Using the same metrics described above, the algorithm's video-level reliability was ICC = .83 and its frame-level reliability was $S = .71$. Second, we measured the algorithm's performance in a publicly-available database

TABLE I

METHODOLOGICAL DETAILS AND SOCIOCULTURAL VARIABLES FOR EACH INCLUDED COUNTRY

Country	Participants	Stimuli	Obs. Smi.	Urban.	Afflu.	Pop. Den.	His. Het.	Indiv.	Eth. Fra.
Argentina	9,178	104	7.69	.92	22,600	15.9	37	46	.26
Bangladesh	4,667	36	5.22	.34	3,600	1,298.0	2	20	.05
Brazil	25,795	313	7.75	.86	15,600	24.4	25	38	.54
Chile	2,698	35	8.39	.90	23,500	23.5	35	23	.19
China	135,111	1,024	3.37	.56	14,100	146.6	1	20	.15
Colombia	7,771	107	5.72	.76	13,800	45.0	23	13	.60
Ecuador	4,538	41	6.95	.64	11,300	57.3	23	8	.66
Egypt	13,018	91	10.89	.43	11,800	88.9	2	25	.18
El Salvador	3,545	54	8.46	.67	8,300	296.4	2	19	.20
Honduras	1,963	42	5.77	.55	4,900	78.2	23	20	.19
Hong Kong	4,768	51	4.12	1.00	56,700	6,655.3	3	25	.06
India	332,673	2,425	2.35	.33	6,200	421.0	3	48	.42
Indonesia	82,118	797	1.68	.54	11,100	141.3	2	14	.74
Japan	6,874	32	2.78	.94	38,100	348.2	1	46	.01
Kenya	6,918	70	6.00	.26	3,200	80.7	4	25	.86
Malaysia	7,465	135	6.13	.75	26,300	92.8	5	26	.59
Mexico	220	86	4.23	.79	17,500	62.6	25	30	.54
Nigeria	11,980	94	4.58	.48	6,100	218.6	3	30	.85
Pakistan	19,624	129	7.81	.39	5,000	258.3	3	14	.71
Panama	4,227	54	9.26	.67	21,800	49.2	37	11	.55
Peru	7,348	59	6.14	.79	12,200	23.8	25	16	.66
Philippines	39,025	419	4.62	.44	7,300	338.7	1	32	.25
Russia	14,079	243	6.47	.74	25,400	8.7	5	39	.39
Singapore	2,244	33	4.92	1.00	85,300	8,259.8	10	20	.39
South Africa	18,877	232	4.53	.65	13,200	44.2	8	65	.75
South Korea	3,850	44	4.53	.83	36,500	506.8	1	18	.00
Taiwan	7,420	55	5.74	.77	46,800	725.8	2	17	.27
Thailand	49,272	540	3.27	.50	16,100	133.1	2	20	.63
United Kingdom	484	88	9.34	.83	41,200	264.9	25	89	.12
USA	897	60	12.53	.82	55,800	35.1	83	91	.49
Vietnam	38,079	382	5.75	.34	6,000	304.3	2	20	.24

Obs. Smi. = Observed Smiling Urban. = Urbanization Afflu. = Affluence Pop. Den. = Population Density
His. Het. = Historical Heterogeneity Indiv. = Individualism Eth. Fra. = Ethnic Fractionalization

(i.e., AM-FED; [38]). The area under the ROC curve when tested on every frame from this database was $A' = 0.94$.

C. Sociocultural Variables

On the basis of previous research and the availability of measures for all countries in the current study, we selected six sociocultural variables to describe each country. These were urbanization, affluence, population density, historical heterogeneity, individualism, and ethnic fractionalization.

As a measure of urbanization, we used the percentage of each country's total population living in urban areas, as defined by each country. These numbers were accessed via [39]. Because an urbanization score for Taiwan was not available through [39], this score was added from [40].

As a measure of affluence, we used each country's gross domestic product (i.e., the value of all final goods and services produced within a country in a given year) divided by its total population [39]. We used gross domestic product at 'purchasing power parity' (i.e., how much all final goods and services produced could purchase in US dollars).

As a measure of population density, we used each country's total population divided by its land area (excluding inland water bodies, in km^2) [39]. As populations are not evenly distributed throughout land areas, this measure should be considered a rough estimate of population density.

As a measure of historical immigration heterogeneity [17], we used the number of countries that have contributed to each country's present-day population since A.D. 1500. These numbers were derived from [41], which was constructed on the basis of textual and genetic data.

As a measure of individualism, we used each country's score on Hofstede's individualism–collectivism index [15]. Higher scores indicate higher individualism. All scores were accessed via the Hofstede Centre website [42].

As a measure of ethnic fractionalization, we used the probability that two randomly selected individuals from each country belong to different ethnic groups [43]. The classification of groups reflects the judgment of ethnologists and is based on racial and linguistic characteristics. Whereas historical heterogeneity captures the diversity of long-history migration, fractionalization captures present-day diversity.

D. Statistical Analyses

In order to quantify each country's level of observed smiling behavior, we first calculated each participant's base rate of smiling during each ad stimulus (i.e., the proportion of video frames during which the participant smiled). We then used each participant's largest base rate as a measure of his or her maximal expressiveness. These measures were then averaged within each country to create the "observed smiling" variable used in subsequent analyses.

We calculated bivariate correlations between all variables

TABLE II

BIVARIATE CORRELATIONS BETWEEN COUNTRY-LEVEL VARIABLES

	1	2	3	4	5	6
1. Observed Smiling						
2. Urbanization	.14					
3. Affluence	.10	.74				
4. Pop. Density[a]	−.41	−.01	.43			
5. Heterogeneity[a]	.56	.43	.17	−.53		
6. Individualism	.40	.22	.29	−.17	.31	
7. Fractionalization	.02	−.28	−.31	−.38	.27	−.09

Note. $n = 31$. [a] This variable was log-transformed.

TABLE III

STANDARDIZED COEFFICIENTS FROM REGRESSION MODEL

	β	95 % CI	p
Observed Smiling			
Urbanization	−0.53	[−1.14, 0.07]	.082
Affluence	0.46	[−0.21, 1.14]	.166
Pop. Density[a]	−0.42	[−0.95, 0.11]	.113
Heterogeneity[a]	0.52	[0.07, 0.96]	.025*
Individualism	0.13	[−0.23, 0.49]	.465
Fractionalization	−0.27	[−0.65, 0.10]	.145

Note. $n = 31$. [a] This variable was log-transformed. * $p < .05$.

and used multiple linear regression to predict observed smiling from the included sociocultural variables. Regression diagnostics [44] were used to identify problems (e.g., non-normality). Because the historical heterogeneity and population density variables were positively skewed, these variables were log-transformed (using base e) prior to analysis.

III. RESULTS

A. Correlation Results

Three variables explained more than $10\,\%$ of the variance in observed smiling ($r^2 > .10$): historical heterogeneity and individualism predicted more smiling, whereas population density predicted less smiling. Table II provides the bivariate correlation for each pairwise combination of study variables.

B. Regression Results

The regression model including all predictor variables explained $47.6\,\%$ of the variance in mean observed smiling. In this model, only the unique contribution of historical heterogeneity was significant (Table III and Figure 2). The standardized regression coefficient (i.e., β) for historical heterogeneity indicates that, when controlling for all other predictor variables, a standard deviation increase in log-transformed historical heterogeneity is associated with a 0.52 standard deviation increase in smiling. In the variables' original units, this result can be interpreted as indicating that, holding all other variables constant, participants from a country with twice the historical heterogeneity as another country would smile during an additional $1\,\%$ of the ad stimulus. To put this in perspective, the vast majority of smiling base rates were between 3 and $9\,\%$; therefore, the difference between two countries on opposite ends of the historical heterogeneity distribution would be considerable.

C. Regression Diagnostics

Ordinary least squares regression makes assumptions that, if violated, can lead to spurious inferences being drawn. Regression diagnostic procedures [44] were used to test these assumptions and to test for outliers and multicollinearity.

The assumption of linearity was supported by a significant F test ($p = .01$) and by visual inspection of scatterplots, the assumption of normality was supported by a non-significant Shapiro-Wilk test ($p = .24$), the assumption of independence was supported by the Durbin-Watson test ($d = 1.53$), and the assumption of homoscedasticity was supported by a non-significant Breusch-Pagan test ($p = .34$).

Two countries (i.e., Egypt and India) were identified as outliers using Cook's Distance measure, which combines information on studentized deleted residuals and leverage. Regression models were estimated both including and excluding these countries; as the pattern of results (i.e., statistical significance) was the same in both models, we only present the model including all countries. Finally, a problematic level of multicollinearity was not evident; specifically, no tolerance values fell below the standard cutoff of 0.10 and no variance inflation factors exceeded the standard cutoff of 10.00.

IV. DISCUSSION

In the current study, we explored the relationship between culture and facial behavior using large-scale observational analyses. Specifically, we used several sociocultural variables to try to predict the average amount of smiling shown by participants in 31 countries. This approach enabled us to test five hypotheses inspired by theory and previous research.

Hypothesis 1 was that different countries would have very different levels of observed smiling. As shown in Figure 2, there was considerable spread in the amount of smiling that was observed in different countries ($M = 6.16, SD = 2.50$). Observed smiling ranged from a minimum of $1.68\,\%$ in Indonesia to a maximum of $12.53\,\%$ in the USA; this means that the average American participant smiled more than seven times longer than the average Indonesian participant. Thus, although no country had an average over $15\,\%$, there was still considerable variability in facial behavior to explain.

Hypothesis 2 was that countries higher in individualism would have more observed smiling. This hypothesis was supported by a medium positive correlation between observed smiling and individualism. However, this variable shared considerable variance with other sociocultural variables and its unique contribution to the prediction of observed smiling was not significant. This finding differs somewhat from previous research [17], [27], which found significant effects for individualism even after controlling for other sociocultural variables. This divergence may indicate that individualism is more influential in shaping cultural norms than actual behavior. Alternatively, it may be that smiling is especially influenced by other cultural factors.

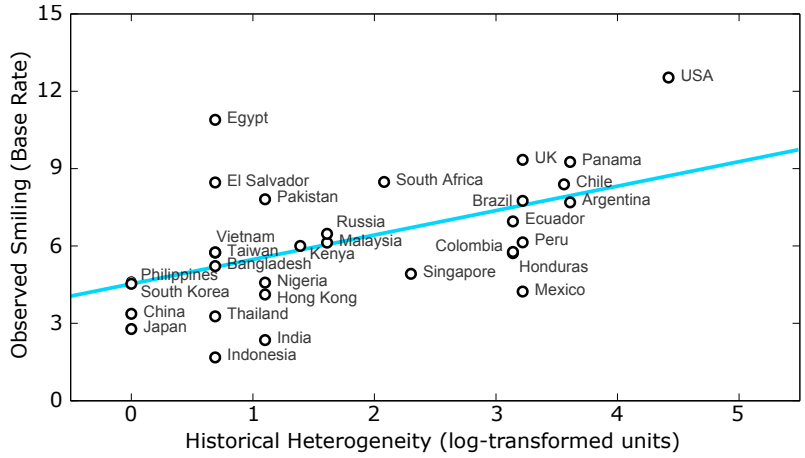

Fig. 2. Scatterplot showing the association between observed smiling and historical heterogeneity.

Hypothesis 3 was that countries with higher historical immigration heterogeneity would have more observed smiling. This hypothesis was supported by a large positive correlation between observed smiling and historical heterogeneity. This variable's unique contribution to the prediction of observed smiling was also significant, despite it sharing considerable variance with other sociocultural variables. These findings (derived, for the first time, from observational measures of actual behavior) constitute an important piece of convergent evidence for the theory that historical heterogeneity explains, and perhaps contributes to the development of, cultural norms for emotional expression and behavior [17].

Hypothesis 4 was that countries with higher affluence would have more observed smiling. This hypothesis was weakly supported by a small positive correlation between observed smiling and affluence, and this variable's unique contribution to the prediction of observed smiling was not significant. These results do not match the findings of Matsumoto et al. [27], who found a larger correlation between expressiveness and affluence, as well as that affluence uniquely predicted expressiveness when controlling for individualism and population density. Numerous differences between the two studies may account for this discrepancy, including those between behaviors (smiles versus any emotion prototype), contexts (market research versus Olympic games), samples (e.g., the specific countries and the number of participants per country), and statistical methods (country-level versus participant-level regressions). We also note that Rychlowska et al. [17] did not find a significant correlation between affluence and self-reported expressivity.

Hypothesis 5 was that countries with higher population density would have more observed smiling. This hypothesis was contradicted by the results. In our sample, there was a medium *negative* correlation between observed smiling and population density, although this variable's unique contribution to the prediction of observed smiling was not significant. In addition to the study differences just listed, this discrepancy may be related to differences in how population density was estimated. Matsumoto et al. [27] divided the current population by the total amount of *arable* land, whereas we used the total amount of land as the denominator. Rychlowska et al. [17] also found a negative association between population density and self-reported expressivity.

Our results provide convergent support for the hypothesis that a long history of cultural diversity is positively associated with norms encouraging nonverbal expressivity. Furthermore, they underscore the predictive power of the historical heterogeneity measure, which accounted for unique variance in observed smiling above and beyond other variables such as individualism and ethnic fractionalization.

Another possible explanation for our results is that cross-cultural differences in the function of smiles interacted with our experimental context. Rychlowska et al. [17] found that smiles are used to signal affiliation in historically heterogeneous societies and to signal dominance in historically homogeneous societies. Given this difference, we may have found greater smiling in historically heterogeneous societies because our experimental context (of viewing television ads in a market research facility) afforded more opportunities for participants to affiliate than to dominate. Perhaps we would have found different results in a different context (e.g., a competitive or conflictual task). Further observational research will be necessary to explore this possibility.

Limitations of the current study largely derive from the market research studies that provided the data. First, relatively little information was collected about individual participants. However, there are likely to be systematic differences within-countries between participants with different demographic characteristics, personality traits, and values. Future research could use multilevel modeling techniques [45] to incorporate such variables while still accounting for the fact that participants from the same country are likely to be more similar than participants from different countries. In an expansion of this study, we plan to use such techniques to control for self-reported participant gender.

Second, relatively little information was collected about individual ad stimuli and, although we made an effort to address this possibility by restricting our sample to countries

with many stimuli, it is possible that the stimuli differed in relevant ways between countries (on average). For instance, some countries may be more likely than others to produce ads that elicit smiles. We suspect that such tendencies would be driven by (or at least correlated with) sociocultural variables that we included (e.g., affluence and individualism), but empirical examination of the ads themselves would be worthwhile. We are currently exploring the possibility of performing content analysis on ads from each country.

Finally, some areas of the world were less represented in the current study than others. In particular, representation was sparse for Europe, Africa, the Middle East, and the continent of Australia. Including additional countries would improve the representativeness and statistical power of our analyses.

Other directions for future research include the comparison of behavior in different social and experimental contexts, the observational measurement of behavior other than smiles, and the refinement of the historical heterogeneity construct. Rychlowska et al. [17] operationalized this construct as the number of source countries contributing to a country's present-day population since A.D. 1500. Because migration events varied in scale and character, more research is needed to account for the timing and size of such migrations, as well as for the degree of cultural 'similarity' between source and destination countries. We also need to better understand the mechanisms underlying the impact of historical heterogeneity on expressive norms and behavior.

Finally, the current study demonstrates the power of automated facial expression analysis to enable large-scale research that would not be feasible using traditional observational methods. Due to the prohibitive cost of manual coding by expert human coders, previous studies on cross-cultural differences in facial behavior had to choose between a high number of countries (e.g., [27] included 84 athletes from 35 countries) or a high number of participants (e.g., [25] included 433 children from 3 countries). However, automated coding is highly scalable and enabled us to include both a high number of countries and a high number of participants (i.e., 866,726 adults from 31 countries). Larger samples are desirable for two reasons: first, they provide more power to detect true effects, and second, they make it more likely that statistically significant results reflect true effects [46].

Large-scale analyses of behavior across (or even within) countries will probably never be possible using traditional methods of data collection and observational measurement; instead, technology-based methods will almost certainly be needed to conduct such research. We regard the development and interdisciplinary application of such methods to be two of the most important contributions that the field of affective computing can make to the scientific community. In this study, and in [29], we have demonstrated that large-scale observational research is feasible. We hope that this type of research will become increasingly common moving forward.

Supplementary Material

Derived data, syntax for statistical analysis, and other information is available from `http://osf.io/4zxms/`.

References

[1] J. A. Russell and J.-M. Fernández-Dols, Eds., *The psychology of facial expression*. New York, NY: Cambridge University Press, 1997.

[2] P. Ekman and E. L. Rosenberg, *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACS)*, 2nd ed. New York, NY: Oxford University Press, 2005.

[3] J.-M. Fernández-Dols and M.-A. Ruiz-Belda, "Spontaneous facial behavior during intense emotional episodes: Artistic truth and optical truth," in *The psychology of facial expression*, J. A. Russell and J.-M. Fernández-Dols, Eds. New York, NY: Cambridge University Press, 1997, pp. 255–274.

[4] L. P. Naumann, S. Vazire, P. J. Rentfrow, and S. D. Gosling, "Personality judgments based on physical appearance," *Personality and Social Psychology Bulletin*, vol. 35, no. 12, pp. 1661–1671, 2009.

[5] M. LaFrance, M. A. Hecht, and E. L. Paluck, "The contingent smile: A meta-analysis of sex differences in smiling," *Psychological Bulletin*, vol. 129, no. 2, pp. 305–334, 2003.

[6] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, and D. P. Rosenwald, "Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses," *Image and Vision Computing*, vol. 32, no. 10, pp. 641–647, 2014.

[7] R. E. Kraut and R. E. Johnston, "Social and emotional messages of smiling: An ethological approach," *Journal of Personality and Social Psychology*, vol. 37, no. 9, p. 1539, 1979.

[8] A. J. Fridlund, "Sociality of solitary smiling: Potentiation by an implicit audience," *Journal of Personality and Social Psychology*, vol. 60, no. 2, pp. 229–240, 1991.

[9] K. Schneider and I. Josephs, "The expressive and communicative functions of preschool children's smiles in an achievement-situation," *Journal of Nonverbal Behavior*, vol. 15, no. 3, pp. 185–198, 1991.

[10] D. Matsumoto, "Culture and nonverbal behavior," *Handbook of nonverbal communication*, pp. 219–236, 2006.

[11] D. Matsumoto, S. H. Yoo, J. Fontaine, A. M. Anguas-Wong, M. Arriola, B. Ataca, ..., and E. Grossi, "Mapping expressive differences around the world: The relationship between emotional display rules and individualism versus collectivism," *Journal of Cross-Cultural Psychology*, vol. 39, no. 1, pp. 55–74, 2008.

[12] E. Suh, E. Diener, S. Oishi, and H. C. Triandis, "The shifting basis of life satisfaction judgments across cultures: Emotions versus norms," *Journal of Personality and Social Psychology*, vol. 74, no. 2, pp. 482–493, 1998.

[13] O. Klineberg, *Social psychology*. New York, NY: Holt, 1940.

[14] P. Ekman and W. V. Friesen, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding," *Semiotica*, vol. 1, pp. 49–98, 1969.

[15] G. H. Hofstede, *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*, 2nd ed. Sage Publications, 2001.

[16] D. Matsumoto, "Cultural similarities and differences in display rules," *Motivation and Emotion*, vol. 14, no. 3, pp. 195–214, 1990.

[17] M. Rychlowska, Y. Miyamoto, D. Matsumoto, U. Hess, E. Gilboa-Schechtman, S. Kamble, H. Muluk, T. Masuda, and P. M. Niedenthal, "Heterogeneity of long-history migration explains cultural differences in reports of emotional expressivity and the functions of smiles," *Proceedings of the National Academy of Sciences*, vol. 112, no. 19, pp. E2429–E2436, 2015.

[18] B. Mesquita and N. H. Frijda, "Cultural variations in emotions: A review," *Psychological Bulletin*, vol. 112, no. 2, pp. 179–204, 1992.

[19] R. T. Boone and R. Buck, "Emotional expressivity and trustworthiness: The role of nonverbal behavior in the evolution of cooperation," *Journal of Nonverbal Behavior*, vol. 27, no. 3, pp. 163–182, 2003.

[20] W. V. Friesen, "Cultural differences in facial expressions in a social situation: An experimental test on the concept of display rules," Doctoral dissertation, University of California San Francisco, 1973.

[21] P. Ekman, W. V. Friesen, and P. Ellsworth, *Emotion in the human face: Guidelines for research and an integration of findings*. New York, NY: Pergamon Press, 1972.

[22] P. H. Waxer, "Video ethology: Television as a data base for cross-cultural studies in nonverbal displays," *Journal of Nonverbal Behavior*, vol. 9, no. 2, pp. 111–120, 1985.

[23] L. A. Camras, H. Oster, J. Campos, R. Campos, T. Ujiie, K. Miyake, L. Wang, and Z. Meng, "Production of emotional facial expressions in European American, Japanese, and Chinese infants," *Developmental Psychology*, vol. 34, no. 4, pp. 616–628, 1998.

[24] L. A. Camras, R. Bakeman, Y. Chen, K. Norris, and T. R. Cain, "Culture, ethnicity, and children's facial expressions: A study of European American, Mainland Chinese, Chinese American, and adopted Chinese girls," *Emotion*, vol. 6, no. 1, pp. 103–114, 2006.

[25] J. Kagan, D. Arcus, N. Snidman, W. Y. Feng, J. Hendler, and S. Greene, "Reactivity in infants: A cross-national comparison," *Developmental Psychology*, vol. 30, no. 3, pp. 342–345, 1994.

[26] B. S. Kisilevsky, S. M. J. Hains, K. Lee, D. W. Muir, F. Xu, G. Fu, Z. Y. Zhao, and R. L. Yang, "The still-face effect in Chinese and Canadian 3- to 6-month-old infants," *Developmental psychology*, vol. 34, no. 4, pp. 629–639, 1998.

[27] D. Matsumoto, B. Willingham, and A. Olide, "Sequential dynamics and culturally-moderated facial expressions of emotion," *Psychological Science*, vol. 20, no. 10, pp. 1269–1274, 2009.

[28] D. Matsumoto and B. Willingham, "The thrill of victory and the agony of defeat: Spontaneous expressions of medal winners of the 2004 Athens Olympic Games," *Journal of Personality and Social Psychology*, vol. 91, no. 3, pp. 568–581, 2006.

[29] D. McDuff, J. M. Girard, and R. El Kaliouby, "Large-scale observational evidence of cross-cultural differences in facial behavior," *Journal of Nonverbal Behavior*, vol. 41, no. 1, pp. 1–19, 2017.

[30] P. M. Niedenthal, M. Mermillod, M. Maringer, and U. Hess, "The Simulation of Smiles (SIMS) model: Embodied simulation and the meaning of facial expression," *The Behavioral and Brain Sciences*, vol. 33, no. 6, pp. 417–433; discussion 433–480, 2010.

[31] M. LaFrance, *Why Smile: The Science Behind Facial Expressions*. WW Norton & Company, 2011.

[32] D. McDuff, "Crowdsourcing affective responses for predicting media effectiveness," Doctoral dissertation, Massachusetts Institute of Technology, 2014.

[33] P. Ekman, W. V. Friesen, and J. Hager, *Facial action coding system: A technique for the measurement of facial movement*. Salt Lake City, UT: Research Nexus, 2002.

[34] K. O. McGraw and S. P. Wong, "Forming inferences about some intraclass correlation coefficients," *Psychological Methods*, vol. 1, no. 1, pp. 30–46, 1996.

[35] R. L. Brennan and D. J. Prediger, "Coefficient Kappa: Some uses, misuses, and alternatives," *Educational and Psychological Measurement*, vol. 41, no. 3, pp. 687–699, 1981.

[36] E. M. Bennett, R. Alpert, and A. C. Goldstein, "Communication through limited response questioning," *The Public Opinion Quarterly*, vol. 18, no. 3, pp. 303–308, 1954.

[37] T. Senechal, D. J. McDuff, and R. el Kaliouby, "Facial Action Unit Detection using Active Learning and an Efficient Non-Linear Kernel Approximation," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*. IEEE, 2015.

[38] D. McDuff, R. Kaliouby, T. Senechal, M. Amr, J. Cohn, and R. Picard, "Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 881–888.

[39] Central Intelligence Agency, "The World Factbook," 2015. [Online]. Available: https://www.cia.gov/library/publications/the-world-factbook

[40] Department of Economic and Social Affairs: Population Division, "World urbanization prospects: The 2014 revision," United Nations, Tech. Rep., 2015.

[41] L. Putterman and D. N. Weil, "Post-1500 population flows and the long run determinants of economic growth and inequality," *The Quarterly Journal of Economics*, vol. 125, no. 4, pp. 1627–1682, 2010.

[42] The Hofstede Centre, "Country comparison tool," 2016. [Online]. Available: https://www.geert-hofstede.com/countries.html

[43] A. Alesina, A. Devleeschauwer, W. Easterly, S. Kurlat, and R. Wacziarg, "Fractionalization," *Journal of Economic Growth*, vol. 8, no. 2, pp. 155–194, 2003.

[44] J. Cohen, P. Cohen, S. G. West, and L. S. Aiken, *Applied multiple regression/correlation analysis for the behavioral sciences*, 3rd ed. New York, NY: Routledge, 2003.

[45] R. H. Heck and S. L. Thomas, *An introduction to multilevel modeling techniques: MLM and SEM approaches using Mplus*, 3rd ed. New York, NY: Routledge, 2015.

[46] K. S. Button, J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, and E. S. J. Robinson, "Power failure: Why small sample size undermines the reliability of neuroscience," *Nature Reviews Neuroscience*, vol. 14, no. 5, pp. 365–376, 2013.