Anxious voice and avoidant language in interaction with a woman wearing an Islamic headscarf: field-experimental evidence from the Paris metro

Alban Lemasson, Manon Toutain, Francesco Madrisotti, Martin Aranguren

Anxious voice and avoidant language in interaction with a woman wearing an Islamic headscarf: field-experimental evidence from France

Alban Lemasson[1,2], Manon Toutain[1], Francesco Madrisotti[3,4], and Martin Aranguren[3,4,5*]

[1] Univ Rennes, Normandie Univ, CNRS, EthoS (Éthologie animale et humaine) - UMR 6552, F-35000 Rennes, France

[2] Institut Universitaire de France

[3] Centre National de la Recherche Scientifique

[4] Université de Paris, URMIS

[5] Sciences Po, OSC (since September 2021)

* Corresponding author: martin.aranguren@cnrs.fr

Keywords: vocal acoustics, prejudice, social interaction, arousal, intimacy, intergroup anxiety

Acknowledgements: The data have been collected as part of the MIDI project, funded by MITI-Centre National de la Recherche Scientifique with a Momentum grant to Martin Aranguren. The funder had no role in the conduct of this research.
We thank Marie Bénédicte Cazeneuve for serving as the confederate, Veronique Biquand for preliminary statistical work, Antoine L'Azou for helping with the literature survey, Maxime Choblet for methodological input, Corentin Monmasson for technical assistance with Praat, and Gaëtan Roisné-Hamelin for acting as the second coder.

Authors contributions. Alban Lemasson conceived the study, contributed methods, performed the measurements, and drafted the first version of the manuscript. Manon Toutain performed the measurements and explored the data. Francesco Madrisotti collected the data. Martin Aranguren designed the experiment, collected the data, performed the statistical analyses, and drafted the final version of the manuscript.

Abstract. In recent years physiological indexes of emotion have made their comeback as indicators of prejudice, but vocal measures have lagged behind. Based on a field experiment involving live interactions in a public place, the study examines intergroup anxiety as it manifests in the voice and in verbal behavior. Two competing predictions are put to test drawing on the "intergroup anxiety model". According to the first prediction, the headscarf should have a positive main effect on anxiety (indexed primarily by vocal cues of arousal) and verbal avoidance (as the reverse of verbal intimacy). According to the second prediction, a cross-over interaction between the headscarf and the sex of the participant should lead to an increase in arousal and a decrease in intimacy among males, but to the opposite response among females. The results confirm the prediction of a positive main effect of the hijab on primary acoustic indexes of arousal and on avoidance as manifested in the pragmatics of language, belying the rival prediction of a cross-over interaction. The study suggest that, in the quest for ecological validity, the voice can be used in a field setting as a direct indicator of bodily activation in intergroup encounters, unobstrusively and economically.

After a relative withdrawal in the 1980s and 1990s (Guglielmi, 1999), in the first two decades of the 21st century physiological indexes of emotion have made their comeback as indicators of prejudice in the context of intergroup relations. In part, this renewed interest reflects the development of new psychophysiological variables (e.g. neuroimaging, Amodio, 2014; Chekroud et al., 2014), or the novel application of existing ones to the phenomenon of prejudice (e.g. the startle eyeblink response, Amodio et al., 2003; Brown et al., 2006; Mahaffey et al., 2005; March & Graham, 2015; Paulus et al., 2019; Phelps et al., 2000; Vanman et al., 2013). But to a significant extent, it also represents the revival of old measures, rejuvenated by a theoretical return of the pendulum to emotion (vs. cognition) as the key component of prejudice, supplemented by gains both in methodological lucidity and technical sophistication. This resurgence, however, has not concerned all pre-existing psychophysiological measures of prejudice to the same extent. Of these, the center stage has been accorded to responses thought to be controlled by the autonomous nervous system such as facial and electrodermal activity, heart rate, or blood pressure (Amodio, 2009; Dambrun et al., 2003; Graves et al., 2005; Greenland et al., 2012; Kiebel et al., 2017; Littleford et al., 2005).

One remarkable absent in the list of resurrected psychophysiological measures of prejudice is the voice. This is particularly surprising when one considers that an impressive amount of evidence has cumulated in the field of emotion indicating that well-defined vocalizations such as laughs or cries and the acoustic characteristics of human speech vehicle information about underlying emotional states (reviewed in Koolagudi & Rao, 2012; Russell et al., 2003). On the plane of theory, this empirical development finds a parallel in the formulation of the "Motivation Structural Rule Theory" (Morton, 1977) and the "Vocal Affect Expression Model" (Scherer, 1986) as plausible conceptualizations of vocal affect signaling. Additionally, compared to placing sensors on participants' head or trunk, or asking them to provide samples of saliva, the apparatus needed to record vocal signals stands out as remarkably unobtrusive, minimizing the potential awkwardness of the data collection procedure.

One possible reason why vocalizations and speech acoustics, as relatively unobtrusive indicators of emotion, have not spilled over from the field of emotion to that of prejudice is that beyond general arousal these cues do not systematically differentiate between positively and negatively valenced affective states (Russell et al., 2003). One way to infer the valence associated with vocally expressed arousal is to supplement measures of emotional vocalizations and voice acoustics with verbal indexes of positive or negative evaluation. This is the strategy that we adopt in the present paper, by treating linguistic and conversational behaviors indicative of intimacy as indexes of positive valence. Intimacy makes reference to the degree of interest in, or openness toward, or liking of, the interaction partner that an actor's behavior communicates (Patterson, 1982). For simplicity, we reduce interest, openness and liking to dimensions of the underlying

variable of positive evaluation, so that intimacy boils down to the degree of positivity towards the interaction partner that an actor's behavior functions to signal (a definition that comes close to the one given to the cognate idea of "immediacy", Mehrabian, 1972; the reduction of various dimensions of evaluation to the sole positivity-negativity continuum follows the lead of Expectancy Violantions Theory, Burgoon, 1978). Simply put, behaviors characterized as intimate signal or express a positive evaluation of and to the interaction partner. It must be kept in mind, however, that even though they function to communicate a valence appraisal situated somewhere in the positivity-negativy continuum, intimacy signals are not the evaluation itself. Human beings are capable of steering their outer expressive behaviors, and so they may deliberately put in the public domain an evaluation that does not have a correlate in their inner feelings. So in using intimacy cues to approximate judgments of valence we do not simply assume a priori that the former necessarily reflect the latter in a one-to-one fashion, as will be further elaborated below.

In this sense, a perhaps more compelling reason why prejudice research has not resuscitated the voice as an indicator of stress, arousal, or emotion, lies in the uncertainties that surround the degree to which the acoustic features of vocalizations are subject to voluntary control. Measuring prejudice on the basis of psychophysiological indexes is notoriously costly. The main motivation to accept that cost is the hope that by circumventing or supplementing other indicators of prejudice that are under voluntary control (chiefly: self-reports), the researcher can gain access to attitudes that respondents would be otherwise unwilling or unable to express (Guglielmi, 1999). The extent to which the voice is able to fulfill this goal is unknown, and to the best of our knowledge the question has not been directly addressed in the available literature.

Still, one may wonder why we ask this from the voice in the first place. The overwhelming majority of the above cited studies focus on a particular group as the target of prejudice: Blacks in the United States. Following the Civil Rights movement in 1960s, the law protects African Americans from various forms of discrimination. Similarly, ordinary morality condemns the overt expression of anti-Black prejudice in the United States. In this particular context, social desirability biases pose a serious problem to the validity of self-reported measures of prejudice against African Americans. It is this particular context that has created the motivation to use indirect measures, including psychophysiological ones, as a "bona fide pipeline" (Fazio et al., 1995) to prejudice.

Rather than assuming that the voice directly reveals a prejudiced attitude that people would otherwise hide, in this study we start from the fact of prejudice and investigate the changes that it induces in vocal behavior, broadly understood to cover the acoustic, lexical, pragmatic, and conversational dimensions of speech, as well as emotional vocalizations. This shift in perspective is justified by the specific group under investigation, namely women who wear the Islamic headscarf or hijab in France. The practice of hijab-wearing is widely disapproved in France, the European

country were opposition to the headscarf is additionally strongest (CNCDH, 2019; Helbling, 2014; Pew Research Center, 2005). Since people do not distinguish well between disapproving a Muslim practice and disapproving the Muslim person who enacts that practice (van der Noll et al., 2018), it follows that French residents generally disapprove, or hold negative views of, hijab-wearing women. The aim of this study is to examine the vocal changes in arousal and intimacy that occur in real-life interactions with a woman who wears the Islamic headscarf or hijab, compared to a control condition in which the same woman appears with uncovered hair. We do not take for granted that these vocal changes directly give expression to the underlying negative view of the Islamic headscarf that is known to prevail in the population. Rather, for reasons that will be elaborated in the next section, we subject this assumption to empirical scrutiny, allowing for the possibility of an intelligible mismatch between (presumed) inner evaluation and (observable) outer behavior.

One lingering concern in the psychophysiological study of prejudice is the artificiality of the stimuli and therefore the ecological validity of the results (Guglielmi, 1999; Mendes et al., 2002). In applications of the startle modification paradigm to prejudice research, for example, the typical study uses photographs of Blacks and Whites as the stimulus. A more realistic setting is to provoke intergroup dyadic interactions in the laboratory (e.g. Amodio, 2009; Greenland et al., Littleford et al., 2005). While moving from photographs (or vignettes, Vanman et al., 2013) to live interactions is undoubtedly a progress in ecological validity, it remains that the encounter takes place in a laboratory and, perhaps more importantly, that the demographic profile of participants tends to limit itself to university students. The same is true of bioacoustics studies which are typically based on intense stress-provoking experiments (Laukka et al., 2008), on known strong correlations between voice quality and self-scored anxiety (Almeida et al., 2014), or on the identification of emotions purposefully enacted by actors (Banse & Scherer, 1996; Juslin & Laukka, 2001). Using live interactions with a hijab-wearing confederate as the stimulus, here we make a further step towards ecological validity both by provoking the interactions in the "natural" context of a public place, and by sampling participants randomly from the wider population.

Predictions

We derive our predictions from the "intergroup anxiety model," (Stephan & Stephan, 1985) which posits that contact with individuals perceived to be members of other groups (or "outgroups") elicits anticipations of negative consequences that provoke anxiety. These undesired consequences may be of various types, such as embarrassment or discomfort due to the awkwardness of the interaction, fear of being exploited, harmed or ridiculed by outgroup members, or apprehension that members of one's own group will disapprove the contact with the other group. The model identifies a set of antecedents and consequents of intergroup anxiety thus depicted. The antecedents include prior

contact with, and cognitions about, the other group, as well as those features of the situation that provide the immediate context for the intergroup encounter. The consequents cover behavior, cognition and evaluation.

In the present application of the model, we manipulate the antecedent that is concerned with prior knowledge of the other group. As said, the Islamic headscarf is met with disapproval in France; it also represents a highly salient and easily recognizable religious sign. Thus, by putting participants to interact with a confederate who wears the garb in the treatment condition but not in the control condition, in line with previous research (e.g. Aberson & Haag 2007) we expect to observe a positive main effect of the hijab on anxiety, which we operationalize as physiological arousal or activation as it manifests in vocal changes. In line with this prediction, though measuring not vocal changes but blood pressure, salivary cortisol concentration, skin conductance and zygomaticus activity, previous studies incorporating physiological indexes of intergroup anxiety show on the whole that interacting with an outgroup increases anxiety and decreases positive affect (Amodio, 2009; Greenland et al., 2012; Littleford et al., 2005).

The model further acknowledges two types of analytically distinct consequents, namely avoidance and amplification. It sees avoidance as the most common response to anxiety, on the assumption that the normal function of avoidance is to reduce anxiety. In the present study, avoidance and its antonym, approach, are treated as the meanings of the opposite poles of the intimacy construct, understood as a universal dimension of relational communication in the context of face-to-face encounters (Burgoon & Hale, 1984). The study focuses on verbal behavior as a channel of expression of intimacy or its opposite. If the intergroup anxiety elicited by the hijab motivates avoidance, and avoidance manifests itself in negative intimacy, we predict that the level of verbally expressed intimacy will decrease when the confederate wears the headscarf. Accordingly, a recent field experiment (Aranguren, Madrisotti, & Durmaz-Martins, 2021) documented a negative main effect of outgroup status on intimacy behaviors (e.g. an increase in the probability of showing a disdainful nonverbal gesture).

Aside from avoidance, drawing on drive theory (Hull, 1951), the model posits that the generic effect of the anxiety provoked by intergroup contact is to amplify the individual's habitual response. As anxiety is assumed to rise the level of drive, anxiety is conconmitantly expected to lead to the amplification or exaggeration of the individual's response to the intergroup encounter, which may or may not be primarly governed by social norms informed by a history of mutual relations. The less social norms organize these encounters, either because they are not available or because participants voluntarily choose not to follow them, the more the consequent of anxiety will depend on individual differences in traits and values. In this regard, echoing an earlier unexpected but robust finding from the field of nonverbal behavior (reviewed in Patterson, 1982), an

intringuing sex difference has been observed when the intergroup encounter is weakly constrained by such norms: whereas men's response is to decrease the level of intimacy or friendliness that they communicate through their behavior, women's is to increase it (Littleford et al., 2005). This could be a direct expression of women's weaker attachment to established beliefs and practices (e.g. social dominance orientation, Sidanius, Pratto & Bobo, 1994), which is known to be positively correlated with intergroup anxiety (Blair, Park & Bachelor 2003). But it could also be an indirect effect of impression management.

So we do not start with any strong a priori claims about the voluntary vs. involontary nature of vocal and verbal cues of arousal. If these are assumed to be involuntary, they are better treated as direct indexes of anxiety. But if they are regarded to be under voluntary control, it may be advisable to give them the status of *a consequent* of anxiety, and so to handle them in the same fashion as intimacy indexes on the amplification hypothesis.

Summing up the discussion, if the voice directly indexes anxiety and the consequent of anxiety is avoidance (the opposite of intimacy), we expect a main effect of the headscarf consisting in an increase in arousal and a decrease in intimacy. If the consequent of anxiety is defined as amplification or exaggeration of habit, and the vocal response is regarded as a consequent under such definition, we expect a cross-over interaction effect of the headscarf and the sex of the participant, or at least simple effects that should hold within one sex group but not the other. On these assumptions, the headscarf should have a simple effect among women consisting in a decrease in arousal and an increase in intimacy, and the simple effect among men should be of opposite sign.

Put in schematic form, these are the predictions that we set out to test in relation to the vocal and verbal response to the headscarf. In interaction with a hijab-wearing woman,

Main effect: overall participants will show P1a) more arousal and P1b) less intimacy;
Simple effect among women: regardless of male participants, female participants will show P2a) less arousal and P2b) more intimacy;
Simple effect among men: regardless of female participants, male participants will show P3a) more arousal and P3b) less intimacy;
Cross-over interaction: P4a) arousal will decrease among females but increase among males, and P4b) intimacy will increase among females but decrease among males.

Method

Design

The experiment follows a between-subjects randomized design with roughly balanced proportions of male and female participants in each of the two experimental conditions and across the six metro stations in which interactions were observed. The goal was to collect at least ten samples representing each combination of sex, experimental condition, and station, that is a total of 240 experimental assays.

Stations selection

Stations were selected at random using a set of filters. The first filter consisted in eliminating all the stations in the upper and lower quartiles by number of passengers, which was a convenient way of taking into account the fact that packed and deserted stations would not offer a suitable environment for the experiment. With the stations in the mid quartiles a random list was then created. The second filter involved visiting the stations in the order stipulated by the random list and ascertaining that the platform was assigned to a single direction (not two) and physically arranged in such a way that there was a single entrance (not many) placed on one of the two longitudinal extremes (not in the middle) of the platform.

Sampling

After a pilot study in March 2018, Martin Aranguren and Francesco Madrisotti performed the experiment between May and June of the same year. The CNRS correspondent of the French commission for the protection of privacy and confidentiality CNIL approved the study and the transportation authority RATP gave us formal clearance to conduct the experiment in the metro. The experimenters made five data collection visits within each of the six selected stations. All visits, scheduled at different weekdays within the same station, had a duration of two hours. Of these, the first hour was assigned to one experimental condition and the second hour to the other condition, balancing for the entire experiment the number of times that each condition was placed first or second in chronological order. During the hour devoted to each condition, in order to recruit an equal number of randomly selected male and female passengers, a method for approximating random selection and another one for stratifying the sampling of men and women was employed. Random selection was approximated with a method of systematic selection: during the time period comprised between the departure of the last train and the arrival of the following one, the confederate approached the first passenger who arrived at the platform. The stratification technique consisted in starting with the method of systematic selection regardless of the sex of the passenger, recruiting one passenger (for example, a man), and then reapplying the method of systematic selection but only to passengers of the opposite sex (women). The third passenger was again selected regardless of sex, the fourth by stratifying by sex, and so on. This means that, in stratifying

our sample, the experimenters relied on their own commonsensical understandings of sexual dimorphism to identify passengers as men or women, and not on passengers' self-reported sexual identity. Data collection visits took place on regular weekdays between 12pm and 2pm. In Paris, this is the only period of the working day in which waiting times are in the range of 3-5 minutes (instead of 1-2), maximizing the chances that the confederate will get to complete the script before the incoming train arrives.

Procedure

On a platform of the local metro, a non-immigrant confederate actress approaches the selected passenger asking for help, on the basis of a standardized script. In one experimental condition, she appears with a hijab; in the other, with uncovered hair. The rest of the clothing is identical, as is the script she follows while interacting with the passengers. Being aware of the fact that she is either using the headscarf or not, the confederate is not blind to the experimental condition. To register the conversation with the passenger, the confederate carries a discreetly mounted portable microphone (VT506 Voice Technologies) and an audio recorder (DR-22WL linear PCM recorder of Tascam). Before approaching, she waits until the selected passenger stops walking and stays standing somewhere on the platform. The passenger stands typically in a position that is perpendicular, on the frontal or coronal plane, to the rails. The confederate, carrying a portable metro map, approaches walking parallel to the rails and stops when the tip of her shoe is at a rough 10 cm distance from the passenger's. The result is a side-by-side arrangement in which confederate and participant form an approximate right angle on the frontal plane. The script divides the interaction in two stages involving different verbal contents and body postures. The first stage consists in locating items on a portable map with confederate and passenger side-by-side. In the second stage, the confederate shifts to a close face-to-face position, asking the passenger to estimate the duration of the trip ahead of her. After the passenger's reply, the confederate laments being late for an important appointment, emphasizes that she needs to contact the person she has to meet, but regrets that her cell phone has run out of battery. After the passenger's reply to this indirect request, a researcher intervenes to unmask the plot and inform the passenger that the interaction has been recorded, requesting consent to process the collected image and audio files. The passenger is then invited to answer to a short questionnaire on sociodemographics.

Measurements and outcome variables

The demographic variables that were measured with the questionnaire are age, educational achievement, income, and religion.

The outcomes reported in the present article rely exclusively on the audio recordings collected in the experiment. The outcomes describing helping and involvement behaviors were measured in 2018 from the video files and have been reported elsewhere (Authors, date). The sound files were produced in WAV 16-bit format at a sampling rate of 44.1 kHz per second. Since the hijab represents mainly a visual stimulus, the fact of using only audio files to take the measurements guarantees complete blindness to the experimental condition, as neither the content of the conversations nor the name of the sound files provide any clues to it.

Using these audio files, Manon Toutain and Alban Lemasson performed the measurements in Spring 2019. The observation period of acoustic measurements are single words that were found to be recurrent in the audio files pertaining to passengers from different groups by condition and sex. Using the program PRAAT (Boersma, 2001), acoustic measurements were performed on a corpus of the following frequently occurring (and emotionally neutral) words: “là”, “ligne”, “minute”, “oui”, and “voilà”. The common observation period for all the non-acoustic outcomes is the entire dialog, from the confederate’s opening to the last sentence interpretable as the closure of the exchange. Acoustic measurements were only performed on audio signals in which low background noise permitted measurements of satisfactory quality. Except for speaking time, which was directly measured, all the other non-acoustic outcomes result from ratings. One fourth of the sample was recorded by a second, independent coder, resulting in satisfactory reliability coefficients for all ratings (ordinal and nominal Cohen’s kappas above 0.7).

Indexes of arousal

1) *Speech acoustics*. We operationalize vocally signaled arousal as “tense voice” (Frick, 1985; Juslin & Laukka, 2001; K. R. Scherer, 2003; Sulter & Wit, 1996). The source–filter theory states that vocal signals result from a two-stage production, with the glottal wave generated in the larynx (the source), being subsequently filtered in the supralaryngeal vocal tract (the filter,(Briefer, 2012; Taylor & Reby, 2010) Tense voice is characterized, among others, by shorter vocalizations, by increased fundamental frequency and amplitude and their respective perturbations (“source”), as well as by a rise in the frequency of all formants (“filter”).

Aside from “source” vs. “filter” effects, we further subdivide the analysis into primary outcomes that have a firm basis in the literature as indexes of arousal, activation or stress (Bachorowski & Owren, 1995; Banse & Scherer, 1996; Forsell et al., 2007; Frick, 1985; Giddens et al., 2013; Juslin & Laukka, 2001; Laukka et al., 2008; Özseven et al., 2018; K. Scherer, 1986; K. R. Scherer, 2003), and exploratory outcomes that do not. The resulting categories are as follows: 1a) *source, primary*: fundamental frequency (F_0) in Hz (mean, maximum and minimum), amplitude in dB (mean, maximum and minimum); 1b) *filter, primary*: formants in Hz (first, second and third) ; 1c) *source,*

exploratory: F_0 coefficient of variation, amplitude coefficient of variation; Wiener entropy (a measure of tonality/randomness going from 0 to minus infinity with 0 being a White noise), F_0 disturbance or "jitter" in %, F_0 amplitude disturbance or "shimmer" in %; 1d) *filter, exploratory*: Highest-pitched frequency in Hz, Temporal position of the maximum amplitude in % of the total duration. As a further exploratory index of arousal that falls neither within the source nor the filter family of outcomes, we also considered the duration of the signal.
2) *Emotional vocalizations*. In this category we consider laughter and conventionalized onomatopoeia indicative of surprise (e.g. "ah!", "oh!") or disfluency ("euh" in French(Cook, 1969).

Indexes of intimacy

1) *Lexical intimacy.* The category covers polite words (e.g. "Hello", "Goodbye", "You're welcome") and apology words (e.g. "I'm sorry", "Excuse me", "I beg your pardon").
2) *Pragmatic intimacy*. Included here are utterances that contextually function to encourage or discourage the continuation of the exchange (ter Maat et al., 2010). We call these "dialog openings" and "dialog closures", respectively. Examples of openings are: "Do you want me to call for you?", "Do you want to text someone?", "The batteries of my cell are dead either but a friend of mine is coming and you can use her phone", "We can go together in the train". Here are some examples of closures: "Ask someone else", "I have to go", "My phone doesn't work", "I can't help you", "It's too complicated".
3) *Conversational intimacy*. As conversational outcomes, we consider speaking time (aggregate length of turns at talk) and the attempt to interrupt the interlocutor, assuming that more time spent speaking and less interruptions reflect higher intimacy (Goldberg, 1990).

Statistical analyses

The data were analyzed with hierarchical, multilevel models of the "varying intercepts, varying slopes" type (Gelman & Hill, 2007) estimated with Bayesian inference. This type of model offers an elegant solution to the problem of multiple comparisons posed by the need to assess treatment effects across numerous indexes of arousal and intimacy, given that with each additional test comes an increase in the risk of falsely rejecting the null hypothesis of no treatment effect (Gelman et al., 2012). Please refer to Supplemental Materials, Statistics for details.
Three such models were performed: a linear regression on intimacy outcomes ("intimacy model"), a linear probability model on acoustic outcomes indicative of arousal ("acoustic arousal model"), and a Poisson regression on verbal indexes of arousal ("verbal arousal model"). All scripts and data used for estimating the models are available as Supplemental Materials, Code and Data.

Note on reporting style. It is inherent to Bayesian inference to describe the output from each model, namely parameter values, as intervals (more precisely, as posterior probability distributions) instead of point estimates. The type of interval considered here is known as the "central posterior interval" (Gelman et al., 2013), and provides the equivalent of a two-tailed test. Unless otherwise indicated, the default alpha level of all the reported central posterior intervals is the standard 5%. Sample sizes are not provided in separate tables but incorporated to the Figures that present the parameter estimations in graphical form, facilitating access to the sample size behind every single reported parameter.

Results

Main effects

P1a) Overall participants will show more arousal

Acoustic arousal model. As indexed by primary acoustic indexes, averaging over males and females the hijab increases arousal by [3%, 34%] of one standard deviation, and the most likely gap equals 19% (Fig 1, parameter "primary"). Exploratory outcomes, in contrast, do not confirm this overall increase (Fig 1, "exploratory"; in fact, there is an interaction effect between these outcomes, such that the effect of the hijab on primary outcomes is credibly larger than the corresponding effect on exploratory outcomes). As a result, the estimate that averages over all acoustic outcomes, although most of its probability distribution is positive, does not yield a credible effect of the hijab at alpha=0.05 (Fig 1, "overall").

When outcomes are separated according not only to their status (primary vs. exploratory) but also to their familiy (source vs. filter), further relevant contrasts arise. The overall effect of the hijab on acoustic arousal is particularly pronounced when the outcome is primary and source-wise, and consists in an increase of [12%, 32%] of one sd, with 22% being the most likely value (Fig 1, "primary & source"). That effect remains credible, although of a smaller size, when it comes to primary outcomes that concern not source but filter characteristics (Fig 1, "primary & filter"). In contrast, when attention is drawn to the parameters concerned with the effect of the hijab on exploratory outcomes, the estimates are credible for neither source nor filter indexes of arousal (Fig 1, "secondary & source" and "secondary & filter").

Verbal arousal model. Verbal indexes of arousal, in turn, do not indicate any overall differences between experimental conditions.

[Insert Figure 1 about here]

Figure 1: Acoustic arousal model, differences between hijab and no-hijab conditions.

P1b) Overall participants will show less intimacy

Intimacy model. The hijab credibly decreases by [-19%, -5%] the probability of showing a pragmatic sign of intimacy (Fig 2, "pragmatic"). However, this effect is indistinguishable from zero when lexical indexes of intimacy are considered instead (Fig 2, "lexical"), which explains why the grand mean of all intimacy indexes (i.e. the average of pragmatic and lexical outcomes) does not yield a credible difference between conditions (Fig 2, "overall").

[Insert Figure 2 about here]

Figure 2: Intimacy model, differences between hijab and no-hijab conditions.

Simple effects among women

P2a) Regardless of male participants, female participants will show less arousal

Acoustic arousal model. This prediction is not supported, and the opposite seems to be the case when primary indexes of arousal pertaining to the source family are take into account. For this particular subgroup of acoustic indexes, female participants' response to the hijab is an *increase* in arousal in the order of [2%, 24%] (Fig 1, "females, primary & source").

Verbal arousal model. The prediction is not supported.

P2b) Regardless of male participants, female participants will show more intimacy

Intimacy model. Again, this expectation finds no support in the data whereas the opposite receives partial confirmation. Thus for female participants, the *negative* effect of the hijab appears to be particularly powerful when pragmatic signs of intimacy are considered, as the garb sinks by [-27%, -10%] the probability of showing them in the course of the interaction (Fig 2, "females, pragmatic").

Simple effects among men

P3a) Regardless of female participants, male participants will show more arousal

Acoustic arousal model. This prediction is supported overall (Fig 1, "males, overall") and for most subgroups of acoustic arousal indexes, including exploratory outcomes from the filter family (Fig 1, "males, exploratory & filter"). Averaging over primary and exploratory outcomes, source as well as filter indexes yield credible effects of the hijab among males (Fig 1, "males, source" and "males, filter"). When consideration is restricted to primary outcomes that describe source characteristics (and that is: fundamental frequency, amplitude and formants), the hijab appears to provoke a

remarkably potent effect among male participants, estimated to represent [18%, 44%] of one sd with 31% as the most likely value.

Verbal arousal model. In contrast, the prediction is not supported when counts of verbal indexes of arousal are considered.

P3b) Regardless of female participants, male participants will show less intimacy

Intimacy model. The prediction is not supported by the data.

Cross-over interactions

P4a) Arousal will decrease among females but increase among males.

Acoustic arousal model. The prediction of a *cross-over* interaction is not supported, but condition*sex interactions do arise as credible effects from the model on acoustic indexes. The effect of the hijab on acoustic arousal is larger for males than for females when the outcomes are of the source family and primary, or of the filter family and exploratory (Fig 1, "interaction, primary & source" and "interaction, exploratory & filter").

Verbal arousal model. The model on the counts of indexes of verbal arousal does not yield any credible interactions.

P4b) Intimacy will increase among females but decrease among males.

Intimacy model. Again, the data does not support the predicted cross-over interaction but indicates a credible difference in effect between the sexes. Thus the decrease in pragmatic intimacy precipitated by the hijab is credibly [3%, 23%] larger for female than for male participants (Fig 2, "interaction, pragmatic").

Discussion

Drawing on measures extracted from the recordings of live interactions between participants and a confederate in a public place, the present investigation set out to examine the effect of the Islamic headscarf (as a proxy of a negatively valenced intergroup encounter) on a range of vocal indexes of intimacy and arousal, allowing the effect of the garb to vary according to the sex of the participant. Two sets of competing predictions were put to test. On the one hand, assuming that indexes of arousal directly measure anxiety, and that the consequent of anxiety is avoidance, it was predicted that the headscarf would increase arousal and decrease intimacy (main effect hypothesis). On the other hand, positing that indexes of arousal represent consequents of anxiety in the same capacity as intimacy cues, and that the consequent of anxiety is an amplification of the habitual mode of

response, in response to the hijab arousal was expected to rise and intimacy to fall among men, and the opposite was expected of women (cross-over interaction hypothesis).

Overall, the data supports the expectation of a main effect and belies the cross-over interaction hypothesis. The most unequivocal evidence of a main effect concerns acoustic indexes of arousal labeled as primary and describing "source" characteristics, that is fundamental frequency, vocal amplitude and frequency of formants. When these outcomes are considered, the hijab turns out to increase arousal by [3%, 34%] of one standard deviation, with the most likely value being 19%. The increase in arousal precipitated by the headscarf holds even when participants are separated into sex groups, with independently estimated simple effects arising as credible among male as well as female participants. These within-sex simple effects of same sign, in turn, credibly differ in size between them. The increase in arousal provoked by the hijab is notably strong for men, among whom it represents [18%, 44%] of one standard deviation with the most likely value located at 31%. For women, that increase is in the more modest range [2%, 24%]. Thus, when primary source-wise indexes are under examination, the hijab does not only provoke an overall increase in vocally expressed arousal, but it also interacts with sex to precipitate a credibly larger effect among men than among women. This interaction cannot be explained away by the well-known morphological sex differences that account for males' lower fundamental frequency and higher amplitude, because the data were rescaled into z-scores within sex groups before being fed into the model.

Another important finding is a converse interaction effect showing that women, more than men, respond to the hijab by decreasing the probability of showing a pragmatic behavior expressive of positive intimacy (conversational openings), or equivalently of not showing a behavior expressive of negative intimacy (conversational closures and attempted interruptions). Importantly also, when males' and females' responses to the hijab on these outcomes are averaged over, a main effect emerges as credible. That is, overall the hijab decreases the probability of showing a sign of pragmatic intimacy, but more clearly so among women than among men.

While limited to the mentioned subgroups of outcomes, the study confirms the main effect hypothesis and does not yield support to the competing prediction that women, contrary to men, should exhibit lower arousal and higher intimacy in the headscarf condition. What seems to arise instead is a sex difference in the way in which the predicted response is manifested, with a larger positive effect on vocally expressed arousal among men but a converse larger negative effect on pragmatic intimacy among women.

Acoustic outcomes indicative of arousal, it may be recalled, were subdivided into "primary" and "exploratory." Despite the sex difference in intensity, the behavior of primary acoustic outcomes (fundamental frequency, amplitude, frequency of formants) is all in all consistent.

Exploratory outcomes, in contrast, deviate more from the baseline acoustic effect, especially within the group of female participants where they are more likely to take on negative (not positive) values. Last, averaging over primary and secondary outcomes, in the main variables from the "source" family behave similarly to those from the "filter" family, both among women and men.

To sum up, the study confirms the prediction of a main effect of the hijab on (intergroup) anxiety as indexed by primary acoustic outcomes and on avoidance as evidenced in the pragmatics of language.

Conclusion

We close this article with an optimistic message for researchers interested in intergroup anxiety, stress or related phenomena who consider to venture into the field and/or to use unobtrusive recordings of human voices to extract indexes of physiological activation. We did not start the present research with the assumption that vocal indexes of arousal directly measure anxiety, allowing for the possibility that participants may deliberately manage those vocal indexes with a view to influencing the other's impression of themselves. If that were the case, vocal indexes of arousal would be measuring a consequent of anxiety rather than the physiological correlate of anxiety itself (that is, bodily activation).

By supporting the prediction of a main effect of the headscarf on anxiety and avoidance and discarding the competing prediction of a cross-over interaction depending on participant sex, the analyses reported above discredit the voluntary-control view and uphold the original assumption that vocal indexes of arousal directly index physiological activation. If this observation could be generalized, the voice would be offering an effective window into physiology that, at least in certain circumstances, could help to make the measurement of bodily activation in studies of intergroup relations less logistically taxing and more ecologically valid.

References

Aberson, C. L., & Haag, S. C. (2007). Contact, Perspective Taking, and Anxiety as Predictors of Stereotype Endorsement, Explicit Attitudes, and Implicit Attitudes. Group Processes & Intergroup Relations, 10(2), 179–201. https://doi.org/10.1177/1368430207074726

Almeida, L. N. A., Lopes, L. W., Costa, D. B. da, Silva, E. G., Cunha, G. M. S. da, Almeida, A. A. F. de, Almeida, L. N. A., Lopes, L. W., Costa, D. B. da, Silva, E. G., Cunha, G. M. S. da, & Almeida, A. A. F. de. (2014). Características vocais e emocionais de professores e não professores com baixa e alta ansiedade. *Audiology - Communication Research*, *19*(2), 179–185. https://doi.org/10.1590/S2317-64312014000200013

Amodio, D. M. (2009). Intergroup anxiety effects on the control of racial stereotypes: A psychoneuroendocrine analysis. *Journal of Experimental Social Psychology*, *45*(1), 60–67. https://doi.org/10.1016/j.jesp.2008.08.009

Amodio, D. M. (2014). The neuroscience of prejudice and stereotyping. *Nature Reviews*

Neuroscience, *15*(10), 670–682. https://doi.org/10.1038/nrn3800

Amodio, D. M., Harmon-Jones, E., & Devine, P. G. (2003). Individual differences in the activation and control of affective race bias as assessed by startle eyeblink response and self-report. - PsycNET. *Journal of Personality and Social Psychology*, *84*(4), 738–753. https://doi.org/10.1037/0022-3514.84.4.738

Aranguren, M., Madrisotti, F., & Durmaz-Martins, E. (2021). Anti-Muslim behavior in everyday interaction: Evidence from a field experiment in Paris. *Journal of Ethnic and Migration Studies.* 10.1080/1369183X.2021.1953378

Aranguren, M., Madrisotti, F., Durmaz-Martins, E., Gerger, G., Wittmann, L., & Méhu, M. (2021). Responses to the islamic headscarf in everyday interactions depend on sex and locale: A field experiment in the metros of Brussels, Paris, and Vienna on helping and involvement behaviors. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0254927

Bachorowski, J. A., & Owren, M. J. (1995). Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context. *Psychological Science*, *6*(4), 219–224. https://doi.org/10.1111/j.1467-9280.1995.tb00596.x

Banse, R., & Scherer, K. (1996). Acoustic profiles in vocal emotion expressoin. *Journal of Personality and Social Psychology*, *70*(3), 614–636. https://doi.org/10.1037/0022-3514.70.3.614

Blair, I. V., Park, B., & Bachelor, J. (2003). Understanding Intergroup Anxiety: Are Some People More Anxious than Others? Group Processes & Intergroup Relations, 6(2), 151–169. https://doi.org/10.1177/1368430203006002002

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

Briefer, E. F. (2012). Vocal expression of emotions in mammals: Mechanisms of production and evidence. *Journal of Zoology*, *288*(1), 1–20. https://doi.org/10.1111/j.1469-7998.2012.00920.x

Brown, L. M., Bradley, M. M., & Lang, P. J. (2006). Affective reactions to pictures of ingroup and outgroup members. *Biological Psychology*, *71*(3), 303–311. https://doi.org/10.1016/j.biopsycho.2005.06.003

Burgoon, J. K. (1978). A Communication Model of Personal Space Violations: Explication and an Initial Test. *Human Communication Research*, *4*(2), 129–142. https://doi.org/10.1111/j.1468-2958.1978.tb00603.x

Burgoon, J. K., & Hale, J. L. (1984). The fundamental topoi of relational communication. *Communication Monographs*, *51*, 193–214.

Chekroud, A. M., Everett, J. A. C., Bridge, H., & Hewstone, M. (2014). A review of neuroimaging studies of race-related prejudice: Does amygdala response reflect threat? *Frontiers in Human Neuroscience*, *8*, 179. https://doi.org/10.3389/fnhum.2014.00179

CNCDH. (2019). *La lutte contre le racisme, l'antisémitisme et la xénophobie: Rapport 2018* (p. 345). Commission Nationale Consultative des Droits de l'Homme; La documentation française. https://www.cncdh.fr/sites/default/files/23072019_version_corrigee_rapport_racisme.pdf

Cook, M. (1969). Anxiety, Speech Disturbances and Speech Rate. *British Journal of Social and Clinical Psychology*, *8*(1), 13–21. https://doi.org/10.1111/j.2044-8260.1969.tb00580.x

Dambrun, M., Desprès, G., & Guimond, S. (2003). On the multifaceted nature of prejudice: Psychophysiological responses to ingroup and outgroup ethnic stimuli. *Current Research in Social Psychology (University of Iowa)*, *8*(14), 12 p.

Ersanilli, E., & Koopmans, R. (2013). *The Six Country Immigrant Integration Comparative Survey (SCIICS)—Technical report*. WZB Discussion Paper SP IV 2013-102.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of Personality and Social Psychology*, *69*(6), 1013–1027. http://dx.doi.org/10.1037/0022-3514.69.6.1013

Forsell, M., Elenius, K., & Laukka, P. (2007). Acoustic correlates of frustration in spontaneous speech. *TMH-QPSR*, *50*(1), 37–40.

Frick, R. W. (1985). Communicating emotion: The role of prosodic features. - PsycNET. *Psychological Bulletin*, *97*(3), 412–429. https://doi.org/10.1037/0033-2909.97.3.412

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). CRC Press.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., Hill, J., & Yajima, M. (2012). Why We (Usually) Don't Have to Worry About Multiple Comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189–211. https://doi.org/10.1080/19345747.2011.618213

Giddens, C. L., Barron, K. W., Byrd-Craven, J., Clark, K. F., & Winter, A. S. (2013). Vocal Indices of Stress: A Review. *Journal of Voice*, *27*(3), 390.e21-390.e29. https://doi.org/10.1016/j.jvoice.2012.12.010

Goldberg, J. A. (1990). Interrupting the discourse on interruptions: An analysis in terms of relationally neutral, power- and rapport-oriented acts. *Journal of Pragmatics*, *14*(6), 883–903. https://doi.org/10.1016/0378-2166(90)90045-F

Graves, R. E., Cassisi, J. E., & Penn, D. L. (2005). Psychophysiological evaluation of stigma towards schizophrenia. *Schizophrenia Research*, *76*(2), 317–327. https://doi.org/10.1016/j.schres.2005.02.003

Greenland, K., Xenias, D., & Maio, G. (2012). Intergroup anxiety from the self and other: Evidence from self-report, physiological effects, and real interactions. *European Journal of Social Psychology*, *42*(2), 150–163. https://doi.org/10.1002/ejsp.867

Guglielmi, R. S. (1999). Psychophysiological Assessment of Prejudice: Past Research, Current Status, and Future Directions. *Personality and Social Psychology Review*, *3*(2), 123–157. https://doi.org/10.1207/s15327957pspr0302_3

Helbling, M. (2014). Opposing muslims and the muslim headscarf in Western Europe. *European Sociological Review*, *30*(2), 242–257. https://doi.org/10.1093/esr/jct038

Hull, C. L. (1951). *Essentials of behavior*. Yale University Press.

Juslin, P. N., & Laukka, P. (2001). Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. - Abstract—Europe PMC. *Emotion*, *1*(4), 381–412. https://doi.org/10.1037/1528-3542.1.4.381

Kiebel, E. M., McFadden, S. L., & Herbstrith, J. C. (2017). Disgusted but not afraid: Feelings toward same-sex kissing reveal subtle homonegativity. *The Journal of Social Psychology*, *157*(3), 263–278. https://doi.org/10.1080/00224545.2016.1184127

Koolagudi, S. G., & Rao, K. S. (2012). Emotion recognition from speech: A review. *International Journal of Speech Technology*, *15*, 99–117. https://doi.org/10.1007/s10772-011-9125-1

Laukka, P., Linnman, C., Åhs, F., Pissiota, A., Frans, Ö., Faria, V., Michelgård, Å., Appel, L., Fredrikson, M., & Furmark, T. (2008). In a Nervous Voice: Acoustic Analysis and Perception of Anxiety in Social Phobics' Speech. *Journal of Nonverbal Behavior*, *32*(4), 195. https://doi.org/10.1007/s10919-008-0055-9

Laukka, P., Neiberg, D., Forsell, M., Karlsson, I., & Elenius, K. (2011). Expression of affect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation. *Computer Speech & Language*, *25*(1), 84–104. https://doi.org/10.1016/j.csl.2010.03.004

Littleford, L. N., Wright, M. O., & Sayoc-Parial, M. (2005). White Students' Intergroup Anxiety During Same-Race and Interracial Interactions: A Multimethod Approach. *Basic and Applied Social Psychology*, *27*(1), 85–94. https://doi.org/10.1207/s15324834basp2701_9

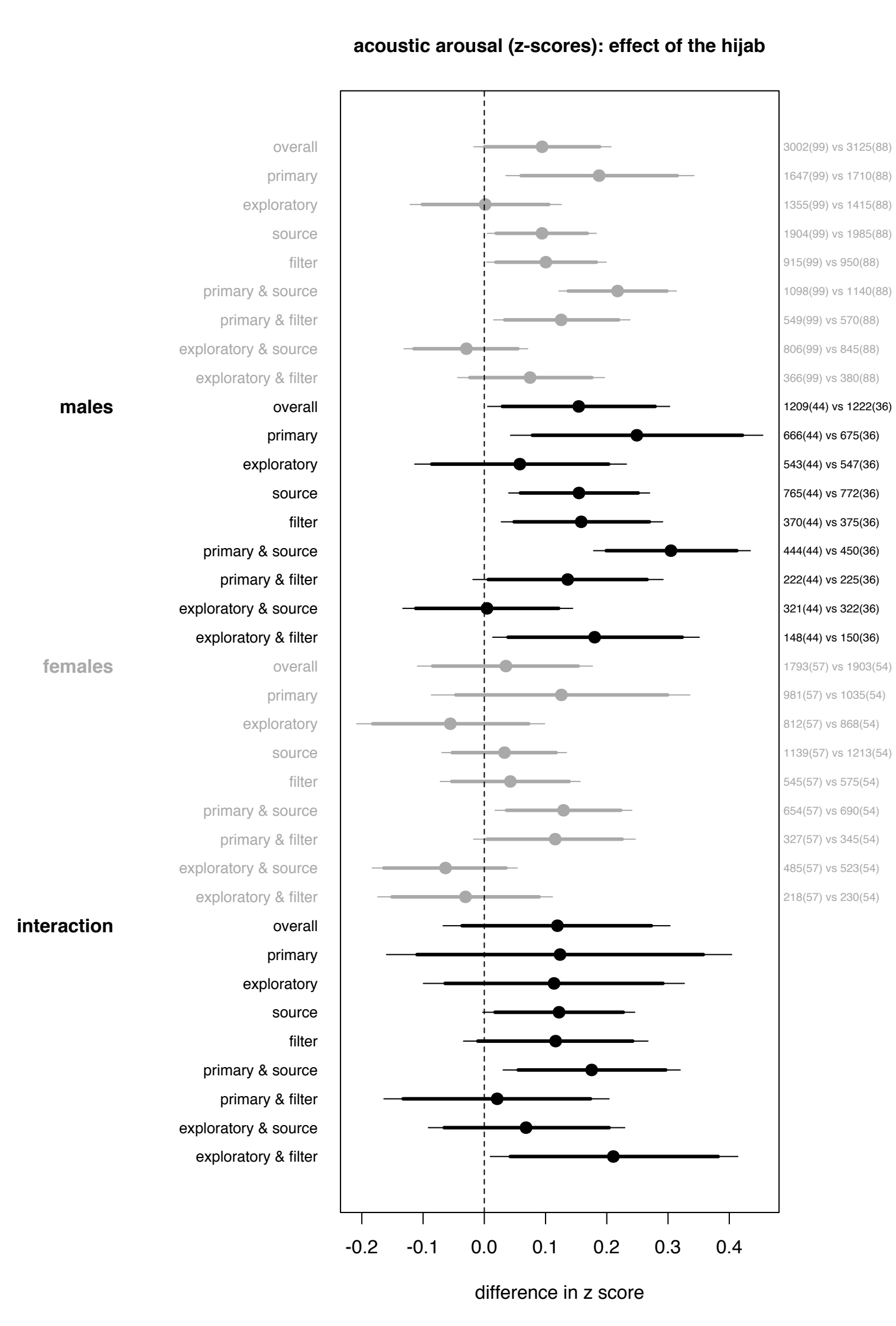
Mahaffey, A. L., Bryan, A., & Hutchison, K. E. (2005). Using Startle Eye Blink to Measure the Affective Component of Antigay Bias. *Basic and Applied Social Psychology*, *27*(1), 37–45. https://doi.org/10.1207/s15324834basp2701_4

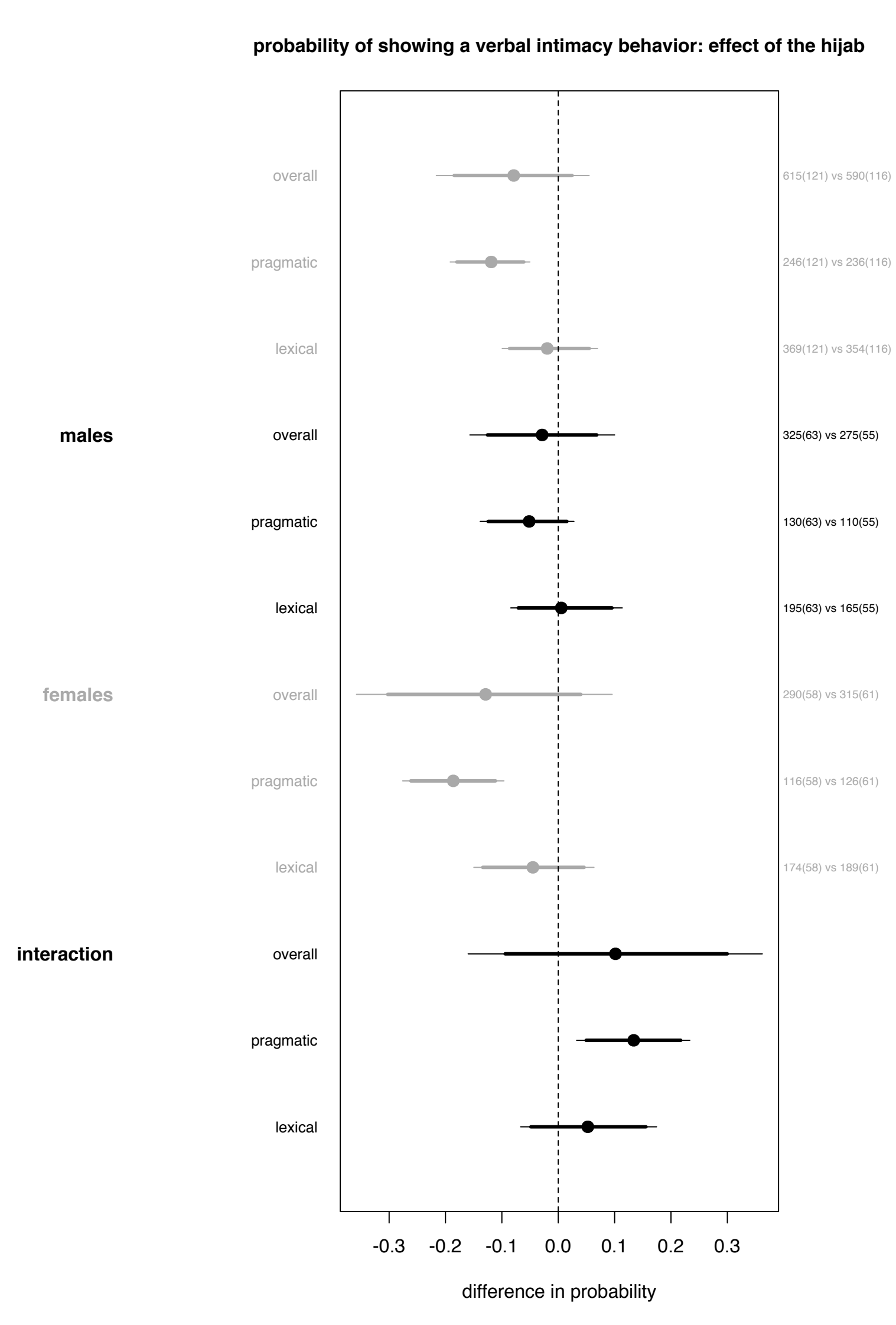
March, D. S., & Graham, R. (2015). Exploring implicit ingroup and outgroup bias toward Hispanics. *Group Processes & Intergroup Relations*, *18*(1), 89–103.

https://doi.org/10.1177/1368430214542256
Mehrabian, A. (1972). *Nonverbal communication*. Aldine-Atherton.
Menahem, R. (1983). La voix et la communication des affects. *L'année Psychologique*, *83*(2), 537–560.
Mendes, W. B., Blascovich, J., Lickel, B., & Hunter, S. (2002). Challenge and Threat During Social Interactions With White and Black Men. *Personality and Social Psychology Bulletin*, *28*(7), 939–952. https://doi.org/10.1177/014616720202800707
Morton, E. S. (1977). On the occurrence and significance of motivation-structural rules in some bird and mammal sounds. *The American Naturalist*, *111*(981), 855–869. https://doi.org/10.1086/283219
Özseven, T., Düğenci, M., Doruk, A., & Kahraman, H. İ. (2018). Voice Traces of Anxiety: Acoustic Parameters Affected by Anxiety Disorder. *Archives of Acoustics*, *43*(4), 625–636. https://doi.org/10.24425/aoa.2018.125156
Patterson, M. L. (1982). A sequential functional model of nonverbal exchange. *Psychological Review*, *89*(3), 231–249. https://doi.org/10.1037/0033-295X.89.3.231
Paulus, A., Renn, K., & Wentura, D. (2019). One plus one is more than two: The interactive influence of group membership and emotional facial expressions on the modulation of the affective startle reflex. *Biological Psychology*, *142*, 140–146. https://doi.org/10.1016/j.biopsycho.2018.12.009
Pew Research Center. (2005). *Islamic extremism: Common concern for muslim and western publics*. https://www.pewresearch.org/global/wp-content/uploads/sites/2/2005/07/Pew-Global-Attitudes-2005-Report-7-14-2005.pdf
Pew Research Center. (2015). *Five facts about the Muslim population in Europe*. http://www.pewresearch.org/fact-tank/2015/11/17/5-facts-about-the-muslim-population-in-europe/
Phelps, E. A., O'Connor, K. J., Cunningham, W. A., Funayama, E. S., Gatenby, J. C., Gore, J. C., & Banaji, M. R. (2000). Performance on Indirect Measures of Race Evaluation Predicts Amygdala Activation. *Journal of Cognitive Neuroscience*, *12*(5), 729–738. https://doi.org/10.1162/089892900562552
Russell, J. A., Bachorowski, J. A., & Fernández-Dols, J. M. (2003). Facial and vocal expressions of emotion. *Annual Review of Psychology*, *54*(1), 329–349.
Scherer, K. (1986). Voice, Stress, and Emotion. In M. H. Appley & R. Trumbull (Eds.), *Dynamics of stress: Physiological, psychological and social perspectives*. Springer.
Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, *40*(1), 227–256.
Sidanius, J., Pratto, F., & Bobo, L. D. (1994). Social dominance orientation and the political psychology of gender: A case of invariance? Journal of Personality and Social Psychology, 67(6), 998–1011.
Stephan, W. G., & Stephan, C. W. (1985). Intergroup anxiety. *Journal of Social Issues*, *41*(3), 157–175. https://doi.org/10.1111/j.1540-4560.1985.tb01134.x
Sulter, A. M., & Wit, H. P. (1996). Glottal volume velocity waveform characteristics in subjects with and without vocal training, related to gender, sound intensity, fundamental frequency, and age. *The Journal of the Acoustical Society of America*, *100*(5), 3360–3373. https://doi.org/10.1121/1.416977
Taylor, A. M., & Reby, D. (2010). The contribution of source–filter theory to mammal vocal communication research. *Journal of Zoology*, *280*(3), 221–236. https://doi.org/10.1111/j.1469-7998.2009.00661.x
ter Maat, M., Truong, K. P., & Heylen, D. (2010). How Turn-Taking Strategies Influence Users' Impressions of an Agent. In J. Allbeck, N. Badler, T. Bickmore, C. Pelachaud, & A. Safonova (Eds.), *Intelligent Virtual Agents* (pp. 441–453). Springer. https://doi.org/10.1007/978-3-642-15892-6_48
van der Noll, J., Saroglou, V., Latour, D., & Dolezal, N. (2018). Western anti-muslim prejudice:

Value conflict or discrimination of persons too? *Political Psychology*, *39*(2), 281–301. https://doi.org/10.1111/pops.12416

Vanman, E. J., Ryan, J. P., Pedersen, W. C., & Ito, T. A. (2013). Probing prejudice with startle eyeblink modification: A marker of attention, emotion, or both? *International Journal of Psychological Research*, *6*, 30–41. https://doi.org/10.21500/20112084.717


acoustic arousal (z-scores): effect of the hijab
overall
primary
exploratory
source
filter
primary & source
primary & filter
exploratory & source
exploratory & filter
males
overall
primary
exploratory
source
filter
primary & source
primary & filter
exploratory & source
exploratory & filter
females
overall
primary
exploratory
source
filter
primary & source
primary & filter
exploratory & source
exploratory & filter
interaction
overall
primary
exploratory
source
filter
primary & source
primary & filter
exploratory & source
exploratory & filter
3002(99) vs 3125(88)
1647(99) vs 1710(88)
1355(99) vs 1415(88)
1904(99) vs 1985(88)
915(99) vs 950(88)
1098(99) vs 1140(88)
549(99) vs 570(88)
806(99) vs 845(88)
366(99) vs 380(88)
1209(44) vs 1222(36)
666(44) vs 675(36)
543(44) vs 547(36)
765(44) vs 772(36)
370(44) vs 375(36)
444(44) vs 450(36)
222(44) vs 225(36)
321(44) vs 322(36)
148(44) vs 150(36)
1793(57) vs 1903(54)
981(57) vs 1035(54)
812(57) vs 868(54)
1139(57) vs 1213(54)
545(57) vs 575(54)
654(57) vs 690(54)
327(57) vs 345(54)
485(57) vs 523(54)
218(57) vs 230(54)
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4
difference in z score


probability of showing a verbal intimacy behavior: effect of the hijab
overall
pragmatic
lexical
males
overall
pragmatic
lexical
females
overall
pragmatic
lexical
interaction
overall
pragmatic
lexical
615(121) vs 590(116)
246(121) vs 236(116)
369(121) vs 354(116)
325(63) vs 275(55)
130(63) vs 110(55)
195(63) vs 165(55)
290(58) vs 315(61)
116(58) vs 126(61)
174(58) vs 189(61)
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
difference in probability

Supplementary Materials, Statistics

The present document spells out the rationale and presents the formulas, as well as other details, of the regression models whose results are reported in the main manuscript.

The data were analyzed with hierarchical, multilevel models of the "varying intercepts, varying slopes" type (Gelman & Hill, 2007) estimated with Bayesian inference. This type of model offers an elegant solution to the problem of multiple comparisons posed by the need to assess treatment effects across numerous indexes of arousal and intimacy, given that with each additional test comes an increase in the risk of falsely rejecting the null hypothesis of no treatment effect (Gelman et al., 2012).

The classical solution to make inferences more conservative in this setting consists in keeping the point estimates stationary (e.g. the mean treatment effect for a given outcome) while making the confidence intervals wider, or equivalently decreasing the critical p value to keep a 5% false alarm rate, as in the well-known Bonferroni correction. Note that in adjustments such as the Bonferroni correction each treatment effect is regarded in isolation from all the others, as though nothing could be learnt about the effect of the hijab on one outcome by considering the effect of hijab on the others.

One alternative well-attuned to the logic of Bayesian inference is to incorporate hierarchy, that is to treat the multiple comparisons (e.g. the treatment effect on outcome 1, on outcome 2, … on outcome *n*) not as totally unconnected but as draws from a common distribution governed by a (yet to be estimated) common mean and variance. In the present application, such assumption makes full sense because every treatment effect has in common with all the others the fact of being the effect of the same treatment, namely the hijab, and of manifesting itself approximately in the same time frame (that is, many outcomes co-occur more or less simultaneously). The consequence of subsuming a collection of parameters (here: approximately simultaneous by-outcome effects of the headscarf) under a common distribution (a procedure termed "partial pooling" because every parameter is modeled yet not independently from the others) is to shift estimates and intervals toward each other by virtue of their common dependence on the parameters of the overarching distribution (these higher-order parameters are known as "hyperparameters"). Hierarchy and partial pooling lead to more conservative estimates because any idiosyncratic estimates (e.g. a large effect of the hijab when all the others are of the same sign but smaller, or a nonzero effect when all the others are nil) will get "shrunk" towards the common higher-order mean, and more or less so

depending on how reliable this individual estimate turns out to be (which is regulated chiefly by sample size). In this manner, partial pooling tends to reduce the number of nonzero treatment effects.

At the same time, in contrast with Null Hypothesis Significance Testing in a traditional frequentist paradigm, Bayesian inference is entirely unaffected by the intentions of the experimenter (Kruschke, 2015). Specifically, a parameter's ability to correctly estimate a treatment effect (e.g. the hijab's) on a given outcome (e.g. vocal fundamental frequency) is completely independent from the experimenter's intention to limit the analysis to that outcome or to go on to estimate treatment effects on other outcomes (e.g vocal amplitude). The output of a model estimated with Bayesian inference, consisting in a set of so-called "posterior distributions", is one and the same whether the analyst considers a single comparison, a set of individual comparisons, or averages over subgroups (or the entirety) of those individual comparisons.

We performed three separate hierarchical models: a linear regression on intimacy outcomes (model 1), another linear regression on acoustic outcomes (model 2), and a Poisson regression on outcomes that concern vocalizations indicative of arousal (model 3). As our goal was to assess not only main effects, but also simple effects within sex groups and also interactions between these, in order to keep the estimates of the treatment effect conservative and reduce the false alarm rate, in models 1 and 3 hierarchy (and therefore "shrinkage") was set to operate *within sex groups*. This means that the estimates for males are unaffected by the estimates for females, although within each sex group the estimate for every outcome is affected by the estimates for all the others. Still, all estimates, whether they are brought under a common distribution or not, are jointly credible – hence the advantage of estimating them together within the same model. Thus, all main, simple and interaction effects of interest can be derived as averages of those by-outcome estimates within sex groups. The slightly more complex model 2 incorporates an additional distinction, namely that between primary and exploratory outcomes. In these models hierarchy was set to operate within sex groups *and within type of outcome* (primary vs. exploratory), in such a manner that the estimates for primary outcomes within a sex group (e.g. fundamental frequency, amplitude and formants for males) are affected by one another, but not by the estimates of exploratory outcomes within the same sex group, and vice-versa. *Mutatis mutandis*, the same considerations as with the simpler models 1 and 3 apply to the calculation of main, simple and interaction effects.

The models to be presented can be described as multi-level, and more precisely as two-level. At one level, the model produces estimates for the data (e.g. the average hijab effect among

males for outcome 1). At a higher-order level, the model produces estimates for the estimates that describe the data (e.g. the average hijab effect among males across all outcomes). The raw data and the R scripts for recoding and rearranging the data, computing the models, summarizing the output in tables and plotting the results are available as Supplementary Materials.

Linear regression on intimacy outcomes

The intimacy indexes were handled within a single linear probability model by first dichotomizing the outcomes. These outcomes had been originally measured as counts (i.e. positive integers from 0 to infinity), but as the regression's goal was to model all intimacy indexes together, the chosen likelihood function could not be the Poisson. More polite words or pragmatic openings, for example, indicate greater intimacy; but more attempted interruptions or pragmatic closures, in contrast, evidence the opposite. It is straightforward to reverse-code indexes of negative intimacy if they are first dichotomized as present vs. absent. The absence of a conversational closure (reverse-coded as 1), for example, indicates more intimacy than its presence (reverse-coded as 0). In contrast, reverse-coding counts (for example, by subtracting every individual count from the maximum count) violates the assumptions of the Poisson model by imposing an arbitrary upper bound (the reverse of the original 0). For ease of calculation, a linear regression (instead of a nonlinear logit or probit one) was used to model the dichotomous intimacy values, considering that when the goal is to test a causal effect (in this case, the difference between the hijab and the control) linear regression on a binary outcome always provides unbiased effect estimates that cannot possibly be out of bounds (Gomila, 2020).

In this model, the units of analysis, *i*, are unique participant-outcome combinations, where the outcome indexes the presence or absence of an intimacy-relevant event in the course of the interaction (presence of polite words, of apology words and of pragmatic openings; absence of pragmatic closures and of attempts to interrupt the confederate). If y_i is the i^{th} dichotomous outcome value, the model states that $y_i \sim \text{normal}(\hat{y}_i, \sigma^2)$, for i = 1…, n=1205, where $\hat{y}_i = X_i\beta^{model1}$. The inputs are the dichotomous outcome y (coded 0 or 1), the participant *j* (237 levels), the outcome *k* (five levels: polite word-yes, apology-yes, opening-yes, closure-no and interruption-no); the participant's age group *l* (two levels), the participant's educational level *m* (two levels), the station where the interaction took place *q* (six levels), whether the participant is a Muslim (1=yes, 0=no), whether the participant is in

the hijab condition (1=yes, 0=no), and whether the participant is a male (1=yes, 0=no). Thus (αs index controls and βs treatment effects),

$$
\begin{aligned}
(1)\ X_i\beta^{model1} = {} & \alpha_0 + \alpha_1 participant_{j[i]} + \alpha_2 outcome_{k[i]} + \alpha_3 ageGroup_{l[i]} + \\
& \alpha_4 eduGroup_{m[i]} + \alpha_5 station_{q[i]} + \alpha_6 muslim * muslim_{[i]} + \\
& \beta_1 hijab.male_{k[i]} * hijab_{[i]} * male_{[i]} + \\
& \beta_2 hijab.female_{k[i]} * hijab_{[i]} * (1 - male_{[i]}), \\
& \text{for } i=1\ldots, n.
\end{aligned}
$$

The predictors concerned with the treatment effects within sex groups, $\beta_1$hijab.male and $\beta_2$hijab.female, are allowed to vary by outcome k, and the resulting by-outcome estimates are given a common distribution with mean and variance to be estimated from the data. More formally, $\beta_1 hijab.male_k \sim normal(\beta_1.hat, \sigma_{\beta 1}^2)$ for 1…, k=5 and similarly $\beta_2 hijab.female_k \sim normal(\beta_2.hat, \sigma_{\beta 2}^2)$ for 1…, k=5. The hyperparameters β_1.hat, $\sigma_{\beta 1}^2$, β_2.hat and $\sigma_{\beta 2}^2$ that describe the common distribution of the by-outcome treatment effects within sex groups are all given noninformative prior distributions (to let the data determine their magnitudes).
The main effect of the hijab is simply the average of $\beta_1 hijab.male_k$ and $\beta_2 hijab.female_k$. Simple effects within sex groups are estimated directly by each of these parameters. Simple effects for lexical vs. pragmatic outcomes, within or across sex groups, are similarly the average of the corresponding subgroups of $\beta_1 hijab.male_k$ and/or $\beta_2 hijab.female_k$ parameters.

Linear regression on acoustic outcomes indicative of arousal

The outcomes concerned with arousal, in turn, had to be separated in two groups, a partition that in practice overlaps the contrast between acoustic and nonacoustic indexes. In principle both types of index could have been analyzed together, but an important difference in the level of measurement advised against this possibility, as acoustic indexes were all continuous but nonacoustic ones discrete counts – with mode equal to zero in two out of three outcomes. Thus the acoustic, continous measurements were analyzed with a linear regression (i.e. a model that assumes that the measurements follow a normal distribution with mean equal to the vector of predictors) whereas the nonacoustic discrete counts were modeled with a Poisson regression (i.e. a model that assumes that the counts follow a Poisson distribution with mean equal to the exponential of the vector of predictors).
The outcomes that were fed into the linear regression on the acoustic outcomes were first put on a common scale by standardizing them as z-scores. Since males and females differ

markedly in average fundamental frequency and vocal amplitude, the z-scores were computed separately for each sex group. The raw measurements did approximate normality for all primary outcomes but not so for all exploratory outcomes, with strong positive skews characterizing the distributions of the frequency coefficient of variation, the duration of vocalization, the dominant frequency, jitter and shimmer. Hence, these acoustic exploratory outcomes were log-transformed before being standardized.

In the linear regression on the acoustic indexes of arousal the units of analysis, *i*, are unique participant-word-outcome combinations (e.g. participant #35, word "là", outcome fundamental frequency; participant #64, word "ligne", outcome shimmer). If y_i is the i^{th} z-score, the model states that $y_i \sim \text{normal}(\hat{y}_i, \sigma^2)$, for i = 1…, n=6341, where $\hat{y}_i = X_i\beta^{model2}$. The inputs are the z-score y, the participant *j* (186 levels), the word *k* (five levels: là, ligne, minute, oui, voilà), the name of the outcome *l* (eleven levels: amplitude, duration, fundamental frequency, fundamental frequency coefficient of variation, highest-pitched frequency, intensity coefficient of variation, jitter, temporal position of the maximum amplitude, shimmer, Wiener entropy); the participant's age group *m* (two levels), the participant's educational level *q* (two levels), the station where the interaction took place *r* (six levels), whether the participant is a Muslim (1=yes, 0=no), whether the participant is in the hijab condition (1=yes, 0=no), whether the participant is a male (1=yes, 0=no) whether the outcome is primary (1=yes, 0=no), the word-outcome combination when the outcome is primary *s* (fifteen levels), and the word-outcome combination when the outcome is nonprimary or exploratory *t* (forty levels). Thus,

$$
\begin{aligned}
(2)\ X_i\beta^{model2} = {} & \alpha_0 + \alpha_1 \text{participant}_{j[i]} + \alpha_2 \text{word}_{k[i]} + \alpha_2 \text{outcome}_{l[i]} + \alpha_3 \text{ageGroup}_{m[i]} + \alpha_4 \text{eduGroup}_{q[i]} + \\
& \alpha_5 \text{station}_{r[i]} + \alpha_6 \text{muslim} * \text{muslim}_{[i]} + \\
& \beta_1 \text{hijab.male.primary}_{s[i]} * \text{hijab}_{[i]} * \text{male}_{[i]} * \text{primary}_{[i]} + \\
& \beta_2 \text{hijab.male.exploratory}_{t[i]} * \text{hijab}_{[i]} * \text{male}_{[i]} * (1 - \text{primary}_{[i]}) + \\
& \beta_3 \text{hijab.female.primary}_{s[i]} * \text{hijab}_{[i]} * (1 - \text{male}_{[i]}) * \text{primary}_{[i]} + \\
& \beta_4 \text{hijab.female.exploratory}_{t[i]} * \text{hijab}_{[i]} * (1 - \text{male}_{[i]}) * (1 - \text{primary}_{[i]}), \\
& \text{for i=1…, n.}
\end{aligned}
$$

βs and hyperparameters as well as main, simple and interaction effects are handled in the same manner as in the previous model. Simple effects regard not only sex groups and primary/exploratory outcomes, but also the distinction between vocal outcomes concerned with "source" vs. "filter" effects.

Poisson regression on verbal indexes of arousal

In this model the units of analysis, *i*, are unique participant-outcome combinations, where the outcome measures the number of times that the participant showed a type of vocalization indicative of arousal (laughter, "ah/oh", "euh"). If y_i is the i^{th} count, the model states that $y_i \sim \text{Poisson}(\lambda_i)$, for i = 1…, n=1205, where $\lambda_i = \exp(X_i\beta^{model1})$. The inputs, vector of predictors, hyperparameter specification and output calculation are the same as in model 1.

Model checks. To approximate the posterior distribution of the parameters of interest we used Markov chain Monte Carlo (MCMC) sampling as implemented by the software Jags (Plummer, 2003) via the programming language R (R Core Team, 2017). We checked that the samples were representative of the posterior distribution through visual examination of trace plots and density plots, on the one hand, and consideration of the Gelman-Rubin statistic of convergence, on the other. None of these checks gave any signs of unrepresentativeness. On the other hand, we checked that the generated samples were large enough (and therefore accurate and stable) by considering a measure called the "effective sample size". The estimates of all the parameters reported below rest on effective sample sizes of over 10,000 (these samples concern the posterior distributions of estimated parameter values, not the data).

References

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gelman, A., Hill, J., & Yajima, M. (2012). Why We (Usually) Don't Have to Worry About Multiple Comparisons. *Journal of Research on Educational Effectiveness*, *5*(2), 189–211. https://doi.org/10.1080/19345747.2011.618213

Gomila, R. (2020). Logistic or linear? Estimating causal effects of experimental treatments on binary outcomes using regression analysis. *Journal of Experimental Psychology: General*. https://doi.org/DOI: 10.1037/xge0000920

Kruschke, J. K. (2015). *Doing Bayesian data analysis*. Academic Press.

Plummer, M. (2003). *JAGS: a program for analysis of Bayesian graphical models using Gibbs sampling*. *124:10*, 1–10.

R Core Team. (2017). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/