Supplemental Material S1
Assessment of Listener Accuracy about Speaker Personality

Introduction

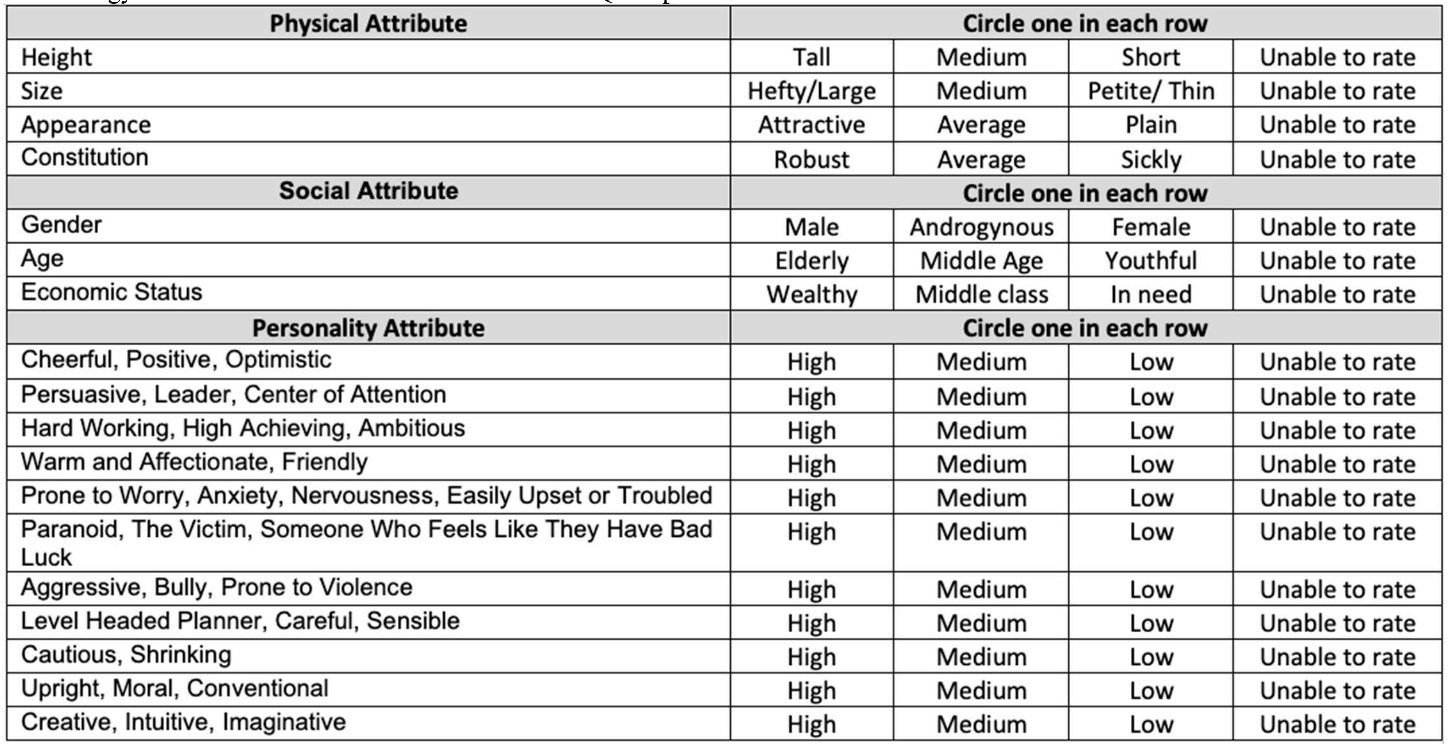
We investigated listener accuracy regarding speakers’ personality traits in addition to listener agreement of speakers’ traits. It is important to understand listener accuracy for a couple of reasons. First, a mismatch between a speaker’s identity and the listener’s perception can result in communication breakdowns. Additionally, an inaccurate assessment by the listener may lead to erroneous conclusions about the speaker, leading to possible social and/or professional consequences. This aspect of our study is unique because previous studies have investigated listener accuracy on stereotyped portrayals of the traits from actors, not the actual personality traits of the (Banse & Scherer, 1996; Scherer, 1978). Other approaches to assess listener accuracy involve disseminating various voice or speech samples to listeners without having obtained any personality information from the speaker, thus preventing comparison of listener ratings to speakers’ personality attributes (e.g., McAleer et al., 2014; Scherer, 1978). In addition, many of these extemporaneous speech investigations do not separate speech and voice production from linguistic content of the speaker, possibly confusing what is said with how it is said.

Methods for assessing listener accuracy

Listeners used the Listener Rating Form (Supplemental Figure S1) to make ratings of each speaker. For Experiments 1 and 2, we calculated accuracy only for the speakers’ personality traits and not the physical or social traits. Speakers’ personality scores were categorized in three levels: high, medium, and low for each primary trait scale of the MPQ-BF using the norms reported in Patrick et al., 2002, and provided by Minnesota Press, Inc. As directed by the official MPQ-BF scoring instructions and norms, trait scores were considered medium scores if speakers fell within one standard deviation (*SD*) of the mean for each scale; they were considered low or high scores if they fell below or above one *SD* of the norm, respectively. If a listener’s rating matched the classification a speaker received on a trait based on the MPQ-BF scoring manual (e.g., a speaker scored low in Social Potency and a listener rated the speaker low in Social Potency), we counted it as an accurate rating. If the listener rating did not correspond to the speaker’s classification on the MPQ-BF, we considered it an inaccurate rating. Blank responses or responses of “unable to rate” did not count against listener accuracy.

Supplemental Figure S1. Listener Rating Form

Listener response form for physical, social, and personality ratings. Personality descriptions for each primary trait scale was taken from terminology used in each the of the 11 scales in the MPQ-BF protocol.



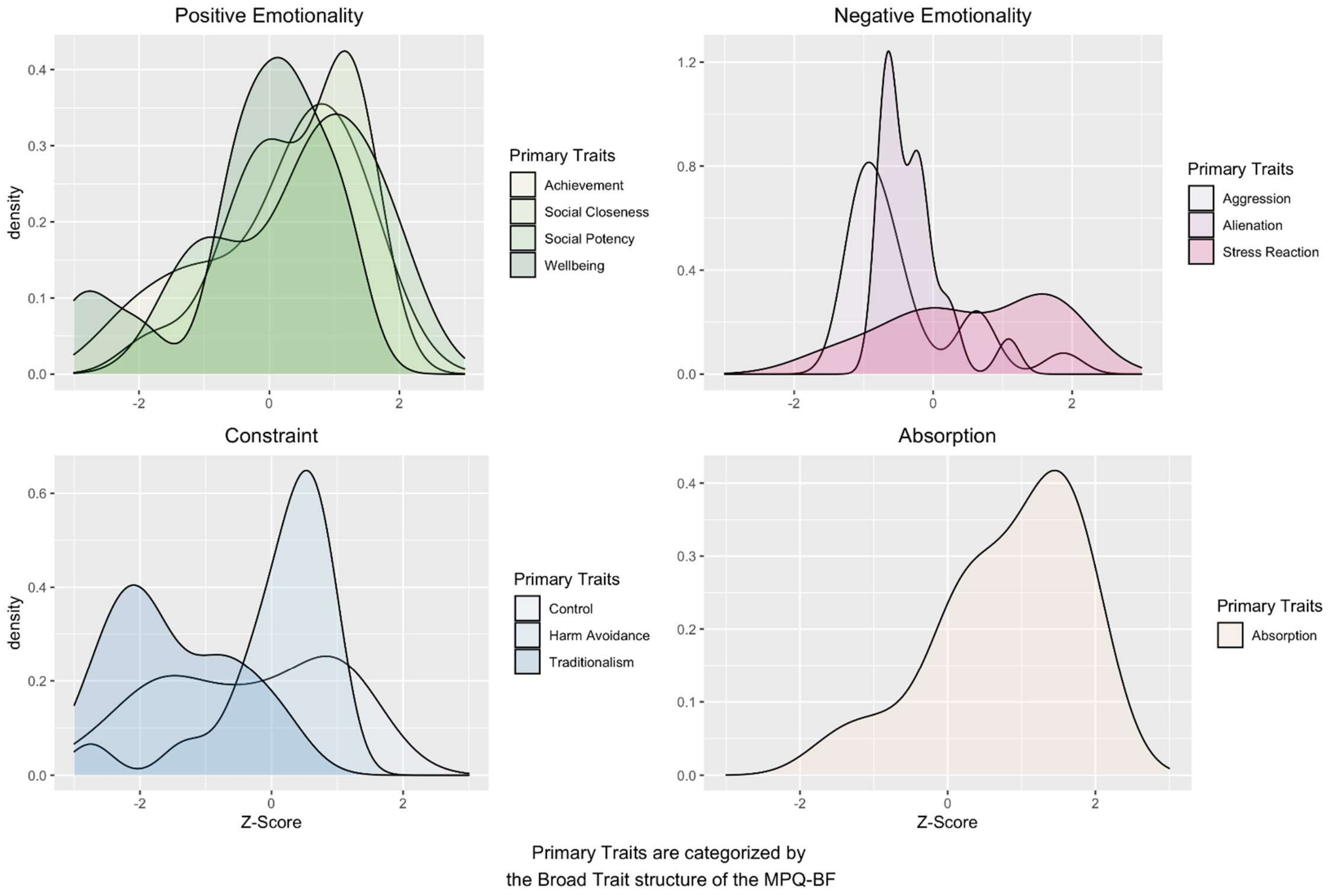
Physical Attribute	Circle one in each row			
Height	Tall	Medium	Short	Unable to rate
Size	Hefty/Large	Medium	Petite/ Thin	Unable to rate
Appearance	Attractive	Average	Plain	Unable to rate
Constitution	Robust	Average	Sickly	Unable to rate
Social Attribute	**Circle one in each row**			
Gender	Male	Androgynous	Female	Unable to rate
Age	Elderly	Middle Age	Youthful	Unable to rate
Economic Status	Wealthy	Middle class	In need	Unable to rate
Personality Attribute	**Circle one in each row**			
Cheerful, Positive, Optimistic	High	Medium	Low	Unable to rate
Persuasive, Leader, Center of Attention	High	Medium	Low	Unable to rate
Hard Working, High Achieving, Ambitious	High	Medium	Low	Unable to rate
Warm and Affectionate, Friendly	High	Medium	Low	Unable to rate
Prone to Worry, Anxiety, Nervousness, Easily Upset or Troubled	High	Medium	Low	Unable to rate
Paranoid, The Victim, Someone Who Feels Like They Have Bad Luck	High	Medium	Low	Unable to rate
Aggressive, Bully, Prone to Violence	High	Medium	Low	Unable to rate
Level Headed Planner, Careful, Sensible	High	Medium	Low	Unable to rate
Cautious, Shrinking	High	Medium	Low	Unable to rate
Upright, Moral, Conventional	High	Medium	Low	Unable to rate
Creative, Intuitive, Imaginative	High	Medium	Low	Unable to rate

Results

Experiment 1

Supplemental Figure S2 shows the distribution of the speakers' MPQ-BF traits. Subjects in the speaker group for Experiment 1 came from a convenience sample. We did not ensure that at least one male and one female speaker were in each of the "low," "medium," and "high" categories.

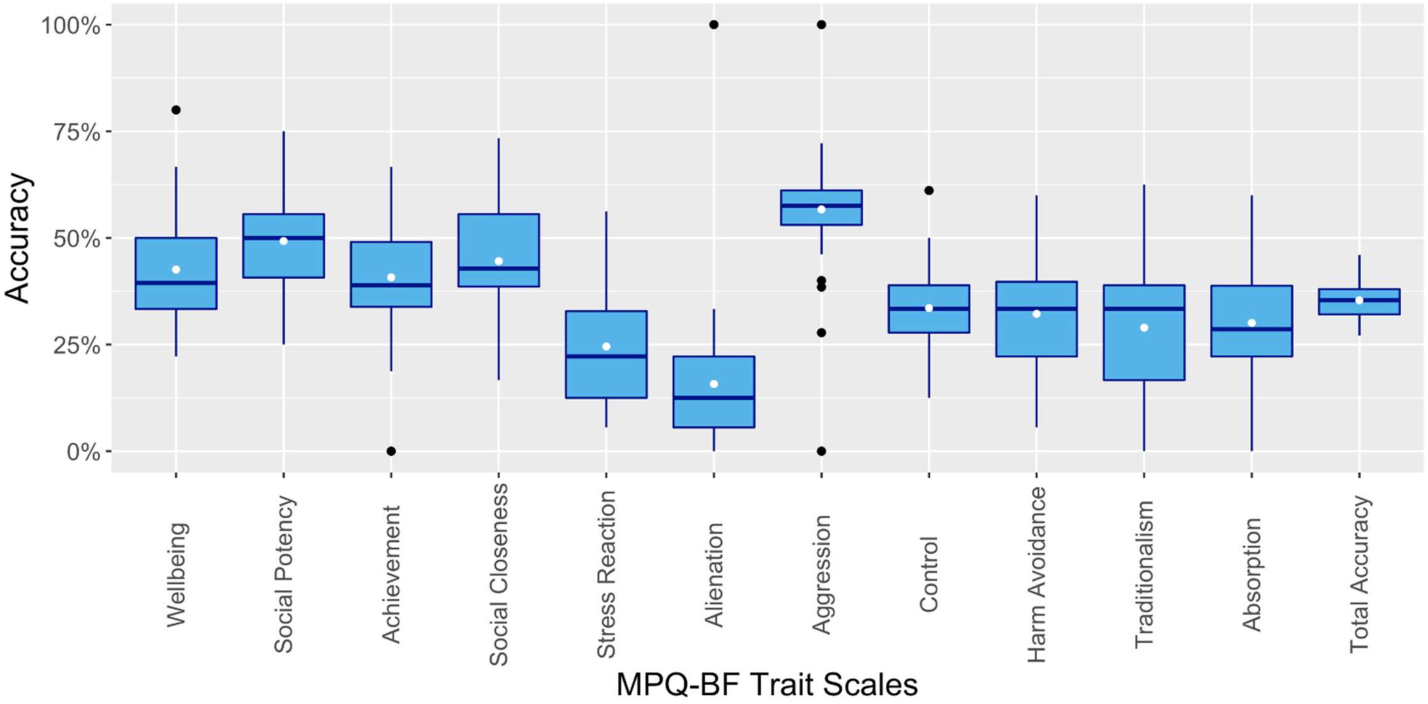
Supplemental Figure S2. MPQ-BF Primary Trait Distribution of Speakers in Experiment 1
The MPQ-BF primary traits are organized by the broad traits they map on to and shown as normalized *z*-scores. *Z*-scores > 1 are classified as "High," *z*-scores between –1 and 1 are classified as "Medium," and *z*-scores < –1 are classified as "Low." Absorption is a trait domain that is distinct from Positive Emotionality, Negative Emotionality, and Constraint. As such, it is displayed separately. The graphs show the relatively likelihood of the *z*-score in the sample, i.e., a density distribution.



Primary Traits are categorized by
the Broad Trait structure of the MPQ-BF

At the group level, listener accuracy was variable across personality traits and speech samples. Mean group-level accuracy ranged from 15% (Alienation) to 56.7% (Aggression). Overall, the accuracy scores were low and tended to fall around the level of random chance (M = 36.3%, SD = 5%). Supplemental Figure S3 shows accuracy data for Experiment 1 across MPQ-BF trait scales.

Supplemental Figure S3. Group Level Listener Accuracy, Experiment 1 - Personality Traits
Group level listener accuracy for Experiment 1, for each of the MPQ-BF primary trait scales, and total accuracy across trait scales. Mean listener accuracy is represented by the white dot on each boxplot. Overall, the accuracy scores were low and tended to fall around the level of random chance (*M* = 36.3%, *SD* = 5%). Additionally, listener responses of "unable to rate" were relatively low (*M* = 7.35%, *SD* = 1.37%). Mean group-level accuracy ranged from 15.7% (Alienation) to 56.7% (Aggression).



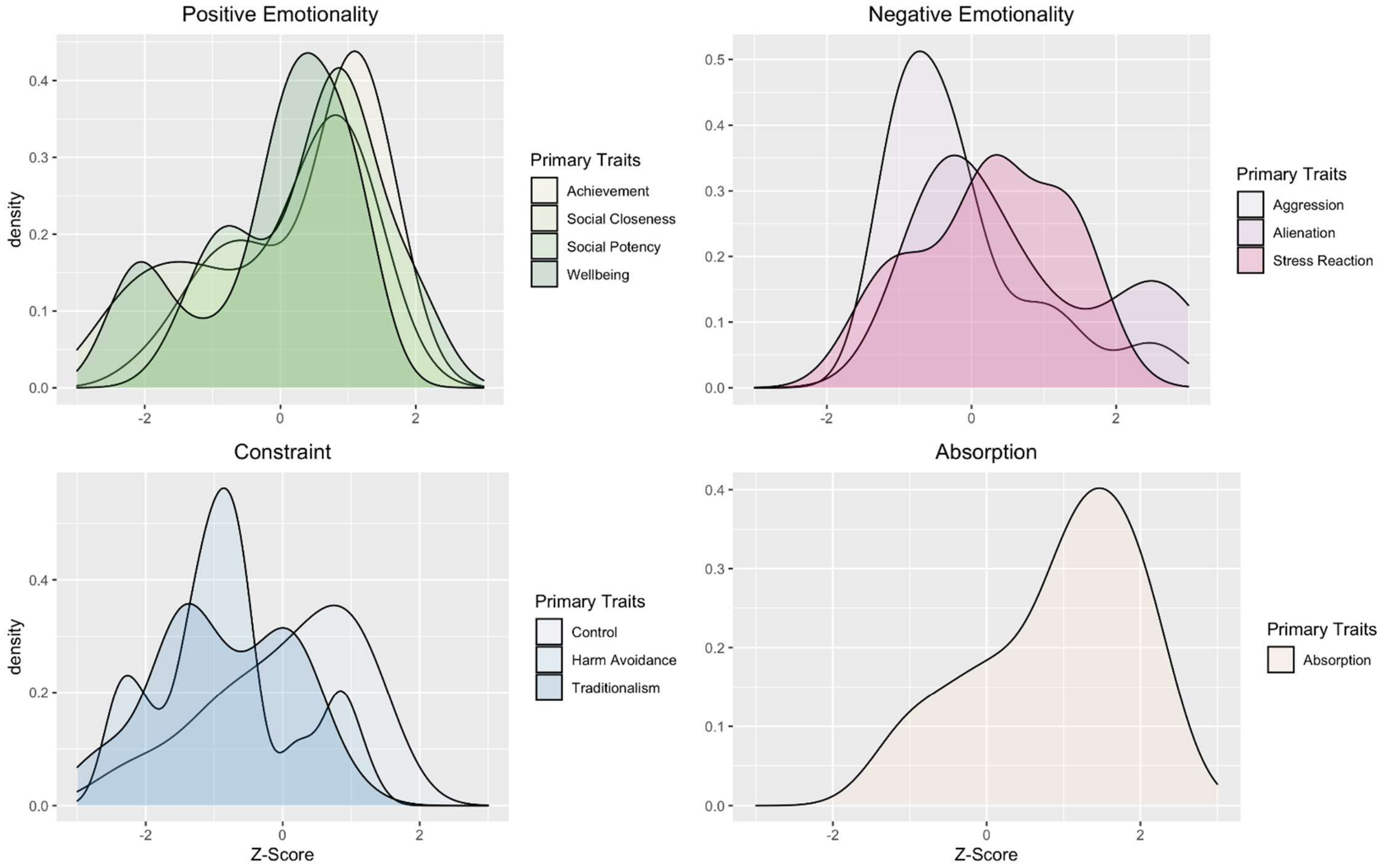
The mean individual listener accuracy across trait scales and speakers was 36.26% (*SD* = 4.9%). However, individual listener accuracy varied greatly. The most accurate listener achieved above-chance levels of accuracy for 9/11 of the scales across all speakers and achieved the highest average accuracy across all trait scales for all speakers (48.21%). Conversely, the least accurate listener only provided above-chance ratings for 4/11 trait scales, with an average accuracy of 30% across all scales for all speakers.

Experiment 2

Supplemental Figure 4 shows the distribution of the speakers' MPQ-BF traits. For Experiment 2, we ensure that each trait was present in the speaker recordings listeners heard, i.e., one male and one female scored "low," "medium," and "high" for each trait.

Supplemental Figure S4. MPQ-BF Primary Trait Distribution of Speakers in Experiment 2

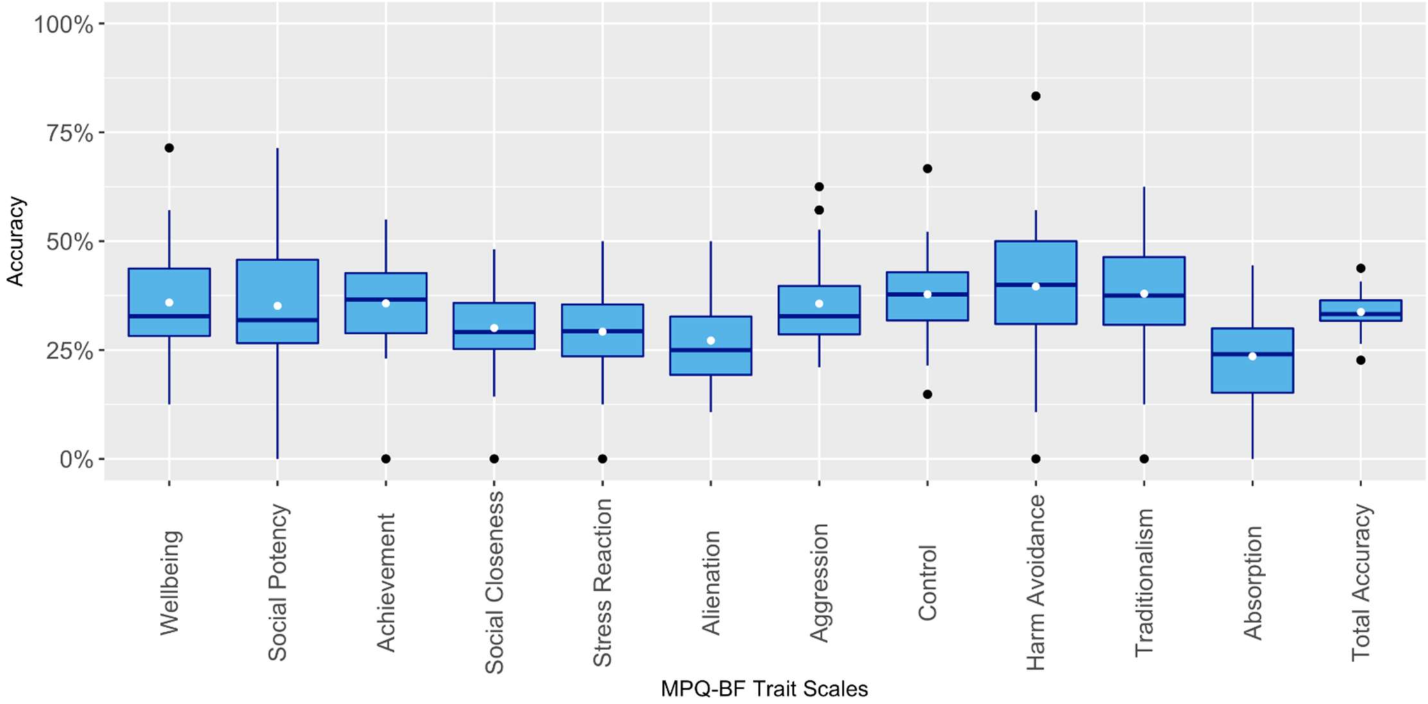
The MPQ-BF primary traits are organized by the broad traits they map on to and shown as normalized *z*-scores. *Z*-scores > 1 are classified as “High,” *z*-scores between –1 and 1 are classified as “Medium,” and *z*-scores < –1 are classified as “Low.” Absorption is a trait domain that is distinct from Positive Emotionality, Negative Emotionality, and Constraint. As such, it is displayed separately. The graphs show the relatively likelihood of the *z*-score in the sample, i.e., a density distribution.



Primary Traits are categorized by
the Broad Trait structure of the MPQ-BF

Group level accuracy varied across personality traits and speech samples. Supplemental Figure S5 shows listener accuracy data for Experiment 2 for each MPQ-BF scale. Average group listener accuracy ranged from 24% (Absorption) to 40% (Harm Avoidance). Overall, the mean group listener accuracy across all trait scales and speakers was roughly equal to chance, i.e., 33% (SD 4%).

Supplemental Figure S5. Group Level Listener Accuracy, Experiment 2 - Personality Traits
Group level listener accuracy for Experiment 2, for each of the MPQ-BF primary trait scales, and total accuracy across trait scales. Mean listener accuracy is represented by the white dot on each boxplot. Average group listener accuracy ranged from 24% (Absorption) to 40% (Harm Avoidance). Overall, the mean group listener accuracy across all trait scales and speakers was roughly equal to chance, i.e., 33% (SD 4%). Similar to Experiment 1, listeners rarely responded "unable to rate" (*M* = 1.8%, *SD* = 1.5%).



As in Experiment 1, the same definition of accuracy was used to evaluate the range of individual listeners' accuracy across traits for all speakers. Experiment 2 yielded similar findings to Experiment 1. While total listener accuracy at the group level was approximately equal to random chance (i.e., 33%), listeners demonstrated variation in their individual accuracy scores across all speakers. The most accurate listener had an average accuracy of 44% for all speakers across all trait scales, and an average accuracy rate that was above 33% for 7/11 scales. Conversely, the least accurate listener had an average score of 22% across all speakers and scales, and only scored above 33% on 2/11 trait scales for all listeners.

Discussion

Regarding accuracy of personality traits, Experiment 1 demonstrated that overall accuracy ratings were relatively low. Certain personality traits appear to be more accurately perceived than others with over half of the listeners accurately rating speakers in Aggression, and nearly half of rating Social Potency with some degree of accuracy. Nonetheless, for most trait scales, listener accuracy remained relatively low, hovering around 33% (i.e., chance). Experiment 2 broadly recapitulated the earlier findings that listeners presented with higher agreement than accuracy. Compared to the cohort of listeners in the profession of communication science and disorders (Experiment 1), listeners in the Experiment 2 cohort (co-ed undergraduate and business school students) were less accurate on 6/11 trait scales of the MPQ-BF.

In addition to limitations outlined in the manuscript, judging accuracy by classifying speakers as "high," "medium," and "low" per the MPQ-BF presents an inherent limitation. Because "medium" is defined as +/– 1 *SD*, fundamentally, 68% of people will fall within this category. However, listeners received "high," "medium," "low," and "unable to rate" as equally weighted rating options. The underlying distribution defining the magnitude of personality traits could have impacted listener accuracy. Future studies should utilize listener rating options that accurately reflect the underlying statistical distribution of traits used to measure speakers.

Based on the findings of these two experiments, it seems that some features (or, more likely, combination of features) of voice and speech render certain personality traits more salient to listeners than other traits. Alternatively, there may be some yet-to-be revealed ability of the listeners themselves (for example, age, experience, professional training, or psychosocial aspects), that facilitate accurate conclusions about the speaker from very little information. Indeed, much evidence shows that personality judgments can be strongly influenced by characteristics of the person making the judgments (Leising et al., 2015; Wessels et al., 2020).

References

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

Leising, D., Scherbaum, S., Locke, K. D., & Zimmermann, J. (2015). A model of "substance" and "evaluation" in person judgments. *Journal of Research in Personality*, *57*, 61–71. https://doi.org/10.1016/j.jrp.2015.04.002

Scherer, K. R. (1978). Personality inference from voice quality: The loud voice of extroversion. *European Journal of Social Psychology*, *8*(4), 467–487. https://doi.org/10.1002/ejsp.2420080405

Wessels, N. M., Zimmermann, J., Biesanz, J. C., & Leising, D. (2020). Differential associations of knowing and liking with accuracy and positivity bias in person perception. *Journal of Personality and Social Psychology*, *118*(1), 149–171. https://doi.org/10.1037/pspp0000218