**Supplemental Materials**

**Voice-Only Communication Enhances Empathic Accuracy**

**by M. W. Kraus, 2017, *American Psychologist***

**http://dx.doi.org/10.1037/amp0000147**

In this supplementary section to the manuscript we report exploratory analyses used to check the robustness of our effects. All materials and data can be found online at <https://osf.io/ux9wa/>.

 **Profile correlations.** Using profile correlations, Experiments 1, 4, and 5 found no significant differences between voice-only communication and combined communication for empathic accuracy: Omnibus *F*(2,296) = 0.67, *p* = .51, Fisher’s LSD *p* = 0.29, *D* = 0.14 (Experiment 1); *t*(188) =0.539, *p* = .59, *DRM* = 0.27 (Experiment 4); Omnibus *F*(3,402) = 36.13, *p* < .001, Fisher’s LSD *p* = .35, *D* = .10 (Experiment 5), however all means were in the predicted direction with voice-only communication providing enhanced empathic accuracy relative to communication across senses. In Experiments 2 and 3 profile correlations revealed significant differences with voice-only communication enhancing empathic accuracy relative to combined communication: *t*(258) = 1.99, *p* = .048, *DRM* = .13 (Experiment 2); Omnibus *F*(2,593) = 19.94, *p* < .001, Fisher’s LSD *p* = .001, *D* = .35 (Experiment 3).

**Gender.** We explored gender differences in empathic accuracy across the five experiments but we did not find systematic differences between groups. In Experiments 1 and 3, men were significantly more accurate than women: *F*(1,294) = 28.38, *p* < .001 (Experiment 1), *F*(1,590) = 6.44, *p* = .01 (Experiment 2). In contrast, Experiments 2, 4, and 5 found that women were more accurate than men, although the effects did not reach statistical significance: *F*(1,262) = 3.65, p = .057 (experiment 2), *F*(1,197) = 3.21, *p* = .075 (experiment 4), *F*(1,398) = 1.33, *p* = .25. The gender of participants only significantly interacted with condition in Experiment 1 F(2,294) = 20.96, *p* < .001 where men were particularly more accurate than women in the face-only condition relative to the other conditions, but this effect did not emerge in Experiment 5 where the same comparison was possible *F*(3,398) = 0.36, *p* = .79. In Experiments 2 and 4 involving live interactions between two strangers we found no differences in empathic accuracy for dyads that were same versus mixed gender *F*(1,262) = 1.51, *p* = .220 (Experiment 2), *F*(1,197) = 1.74, *p* = .19 (Experiment 4). In Experiment 2, gender interacted with same versus mixed gender dyad such that men were least accurate when interacting with other men *F*(1,262) = 5.63, *p* = .034, however this interaction effect between participant gender and dyad gender makeup did not replicate when that comparison was possible again in Experiment 4 *F*(1,197) = 0.30, *p* = .58. Because of the exploratory nature of these analyses and the inconsistency of the findings here, all analyses are reported in the text collapsed across gender.

**Valence.** We also explored condition differences in empathic accuracy as a function of positive and negative affect. For the comparison between voice-only and combined communication we found that positive affect showed significant enhanced accuracy in the voice only condition in three Experiments: *t*(265) = 3.07, *p* = .002 (Experiment 2); LSD *p* = .002 (Experiment 3); *t*(201) = 2.22, *p* = .028 (Experiment 4), and no accuracy differences in two experiments—LSD *p* = .321 (Experiment 1); LSD *p* = .996 (Experiment 5). For negative affect, three Experiments showed significant enhanced accuracy in the voice-only condition compared to the combined condition: LSD *p* = .003 (Experiment 1); *t*(201) = 4.74, *p* < .001 (Experiment 4), LSD *p* = .008 (Experiment 5) and no accuracy differences in two experiments—*t*(265) = 1.36, *p* = .174 (Experiment 2); LSD *p* = .882 (Experiment 3). Overall, these results indicate that enhancement of empathic accuracy in voice-only communication occurred for both positive and negative affect.

**Emotion blends.** We had participants rate emotion blends as well as single items in Experiment 3 to explore whether accuracy findings in Experiment 2 were driven by artifacts related to judging emotions based on blends versus single items. We explored this possibility using a 2 (single item vs. blended emotion rating) X 3 (sensory modality) mixed ANOVA. The analysis revealed no effect of emotion rating type *F*(1,594) = 0.01, *p* = .91, and a significant effect of condition *F*(2,594) = 3.87, *p* = .02 that was consistent with the analysis reported in the results of Experiment 3. Emotion rating type appears not to moderate our empathic accuracy findings. Nevertheless, the different types of emotions items and scales used across experiments in this manuscript is a limitation of the present research (Weidman, Steckler, & Tracy, 2016).

**Emotion expressivity (Experiment 2).** Though the experiment finds some initial evidence for our hypothesis, a couple of alternative explanations are possible. First, the darkened room may have motivated participants to attend more carefully to emotion cues in order to overcome the lack of visual information relative to the lighted room. Second, the darkened room interaction may have heightened the arousal of participants and it was this arousal that led to greater emotion expression and as a result, greater accuracy than in the lighted room. Although we cannot completely rule out these explanations with these data, a few pieces of evidence suggest that these explanations are unlikely to explain our results. In terms of the effort explanation, we examined emotion expressivity differences in the two interactions given its past association with empathic accuracy (Hall, Andrzejewski, & Yopchick, 2009). We measured emotion expressivity in the lighted- and darkened-room interactions by having a team of six trained coders watch the video-recorded interactions and judge the expressivity of each participant on a 4-point Likert scale (0 = *not at all expressive*, 3 = *a great deal of expression*). Coders were instructed to watch the videos and to judge all aspects of verbal and nonverbal emotion expressivity—video recording in the darkened room interaction was possible via our night vision camera. Overall reliability for coded emotion expressions was high (*M* = 1.58, *SD* = 0.84, ɑ’s = .89). If expressivity was heightened in the darkened relative to the lighted room interaction, this would suggest that participants in the darkened room were expending extra effort to elicit understanding of their emotions. Contrary to this prediction, when we analyzed emotion expression as a function of the two room conditions we found a significant effect of room condition *F*(1,132) = 17.31, *p* < .001 in the opposite direction of this effort prediction—participants were more expressive in the light (*M* = 1.68) than they were in the dark (*M* = 1.52). Importantly, this analysis suggests that the effort level of emotion expression did not explain our observed condition differences in empathic accuracy.

**Mean level emotion (Experiments 2 and 4).** In Experiment 2, we examined self-rated mean level affect as a test of the arousal explanation. In this analysis, participant positive (*MDark* = 5.06; *MLight* = 5.01; *t*(265) = 1.33, *p* = .19) and negative (*MDark* = 2.04; *MLight* = 2.10; *t*(265) = -1.17, *p* = .24) affect did not differ as a function of interaction condition, suggesting that the lighted and darkened room interactions engendered similar levels of self-reported affective states. This finding runs counter to the proposition that arousal levels differed by condition.

In Experiment 4, the emotions elicited by the work conversation scenarios appeared to be primarily positive when examining the means. However, examining overall means for positive (*Mvoice* = 3.31, *SDvoice* = 1.90; *Mvoice/visual* = 3.46, *SDvoice/visual* = 1.75) and negative (*Mvoice* = 0.48, *SDvoice* = 0.71; *Mvoice/visual* = 0.56, *SDvoice/visual* = 0.89) emotions revealed no differences in felt emotion as a function of experimental condition (*tpositive*(201) = -1.46, *p* = .15; *tnegative*(201) = -1.31, *p* = .19). This pattern of emotion self-reports is consistent with the pattern observed in Experiment 2—condition differences in empathic accuracy are not driven by mean levels of emotion experience.

**Stereotype use (Experiment 2).** Prior research indicates that people may rely on stereotype beliefs when judging others emotions—that is, what they expect others would feel in a given situation (Lewis, Hodges, Laurent, Srivastava, & Biancarosa, 2012). Based on this research, we assessed “how characteristic” it would be to experience each of the six emotions during discussions of topics related to film and television preferences and food and beverage choices. We recruited 268 workers from Mechanical Turk to answer these stereotype use questions, and respondents made these judgments based on 4-point Likert scales (1 = *uncharacteristic*, 4 = *very characteristic*).

We converted the means from these online stereotype ratings to 7-point Likert scales for each of our six emotions (to correspond to the 7-point rating scale used with the emotion ratings). We then computed an index of stereotype use by calculating the absolute value of the difference between the stereotype use mean and participant estimates of their partner’s emotions during the dark and light room interactions respectively (*M* = 3.12, *SD* = 1.27). When we subjected stereotype accuracy to our analyses, the results revealed no condition differences in stereotype use *F*(1,132) < 1,*ns*, suggesting that participants were not more likely to use stereotypes in the darkened (*M* = 1.32) or lighted (*M* = 1.32) interactions.

**References**

Hall, J. A., Andrzejewski, S. A., & Yopchick, J. E. (2009). Psychosocial correlates of interpersonal sensitivity: A meta-analysis. *Journal of Nonverbal Behavior*, *33*(3), 149-180.

Kenny, D. A., Kashy, D. A., & Cook, W. L. (2006). Dyadic data analysis. *New York, NY: Guilford*.

Lewis, K. L., Hodges, S. D., Laurent, S. M., Srivastava, S., & Biancarosa, G. (2012). Reading Between the Minds The Use of Stereotypes in Empathic Accuracy. *Psychological Science*, *23*(9), 1040-1046.

Weidman, A. C., Steckler, C. M., & Tracy, J. L. (2016). The Jingle and Jangle of Emotion Assessment: Imprecise Measurement, Casual Scale Usage, and Conceptual Fuzziness in Emotion Research. *Emotion.*