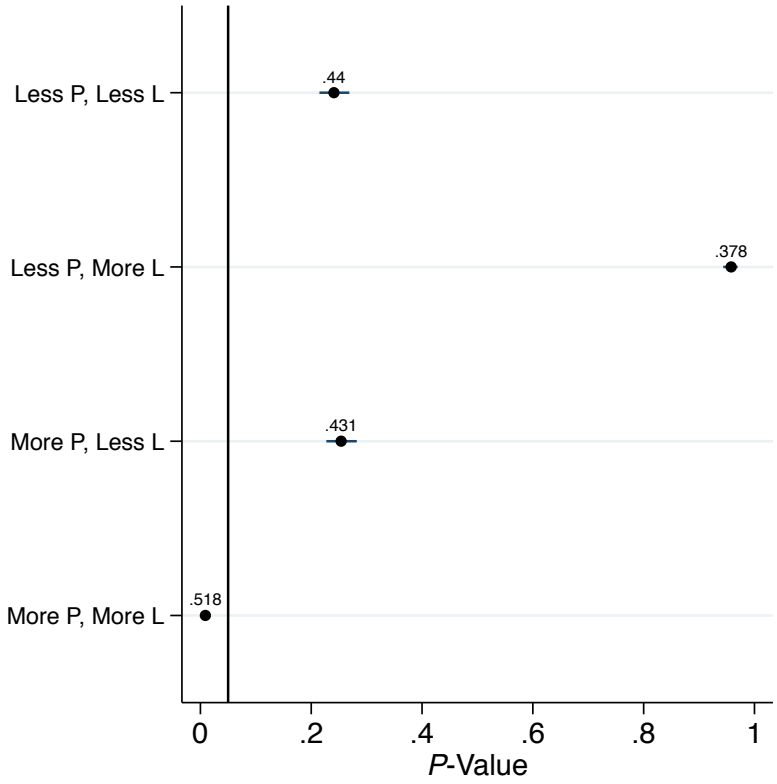

Figure A1. *P*-Plot of Adjusted Predictions

Note: Predicted values are above the markers. Markers themselves are the *p*-values for each predicted value, based on 1,000 Monte Carlo permutation tests. Bars are 95% confidence intervals for the *p*-value. Vertical line is at .05, meaning that any prediction to the left of the line is statistically significant at at least $\alpha = .05$. Predictions are based on Model 3 from Table 4 in the published paper with all controls held at their observed values.

Figure A1 provides an alternative (and perhaps more useful) visualization of the adjusted predictions presented in Figure 1 of the published paper. As mentioned in the paper, Figure 1 does not include confidence intervals because the *p*-values for the effects on which the predictions are based were derived under the randomization model of inference and not the more well-known population model. Figure A1, however, provides the *p*-values for the predicted values themselves (indicated by placement of the marker; the numbers above the markers are the predicted values). The prediction *p*-values were derived in the following steps: (1) a regression model with the observed values is computed and predictions calculated; (2) the values of the dependent variable (the sentiment score) are randomly permuted 1,000 times, and the regression model and subsequent predicted values are re-computed each time; (3) a tally of the number of times the absolute value of the prediction is greater than or equal to the absolute value of the observed prediction is made; and (4) the tally is divided by the number of permutations (in this case, 1,000), providing an empirically-derived *p*-value. Like the *p*-values for the coefficients reported in the paper, a *p*-value

of, say, .04 means that only 4 of the 1,000 instances of that particular prediction had an absolute value greater than or equal to the absolute value of the observed ("real") prediction. Using an $\alpha = .05$, we would reject the null hypothesis that a random data-generating process can account for the observed prediction, since, in fact, an empirically-imposed random process (the random permutations) cannot regularly produce a prediction that is equal to or greater than (the absolute value of) that prediction. We would instead find support for the alternative hypothesis—i.e., that a non-random process generated the data which then lead to that prediction (see Darlington and Hayes 2017:514-16 for more on the difference between "process inference" versus population inference).

Why is Figure A1 useful? For one, it shows that only one combination of perceptibility and legibility levels generates a predicted sentiment that doesn't appear to be due to randomness: the one we hypothesized to be the most conducive to maximizing a journalist's use of sentiment vocabulary, i.e., object-settings where money ideal-typically is both highly perceptible and highly legible. As such, though the substantive difference between the prediction for object-settings where money is both highly perceptible and highly legible and the prediction for object-settings where money is low in perceptibility and low in legibility might seem small ($.52 - .44 = .08$, which is about 44% of a standard deviation in the sentiment variable), only the high-high setting is "statistically significant." So, while one may be tempted to use Wald tests with randomization model *p*-values to compare the high-high prediction with the other three predictions to assess the magnitude of these differences, this doesn't seem to work in the present case. This is because all of the predictions generated except for the high-high one can easily be accounted for by a random data-generating process—i.e., more than 5% of the absolute values of each of those three predictions from the 1,000 permutations are greater than or equal to the absolute value of the observed predictions. For instance, you get a lot of low-low predictions that are greater than .44; so, when you perform Wald tests of this difference with the high-high prediction for each permutation, you can easily get a number of a bunch of small differences or even differences going in the opposite direction. So, while a Wald test of these prediction differences under the randomization model would suggest no significant difference between the high-high and low-low estimates, this not because the two predictions are not *actually* that different from one another. Instead, it is because the absolute value of the high-high prediction is consistently less than the absolute value of the observed prediction (because randomness cannot account for the observed prediction) while the absolute value of the low-low prediction is consistently *greater than* the value of the observed prediction (because randomness *can* account for the observed prediction). It seems to me that the only valid way to use a Wald test of prediction differences under the randomization model of inference is to compare predictions that are statistically significant under the randomization model.

Finally, this plot provides an opportunity for some slight clarification on the interaction between the perceptibility and legibility levels. If we work with a strict $\alpha = .05$ threshold, then the third line of the first paragraph (and leading into the fourth line) on page 34 should technically read: "the coefficient [for the legibility contrast at high levels of perceptibility] becomes both [*marginally*] significant and larger . . . ." This is because the *p*-value for the difference between less legibility and high legibility in high perceptibility settings is .065. However, the marginal significance of the difference between legibility levels does not matter for the argument in the paper. This is because the difference between perceptibility levels is still statistically significant in

high legibility settings (p = .002)—which must be the case, since, as the paper shows, the overall interaction effect is statistically significant and positive. This means that for high perceptibility to be meaningfully different from low perceptibility, the money needs to also exhibit high legibility even though the difference between legibility levels itself is only significant at α = .1. So, high legibility is still required along with high perceptibility to maximize the sentiment score. Figure A1 illustrates that it is indeed the combination of high perceptibility and high legibility that "matters most," since it is only this combination that generates a predicted value that cannot be accounted for by a random data-generating process.

REFERENCE

Darlington, Richard B. and Andrew F. Hayes. 2017. *Regression Analysis and Linear Models: Concepts, Applications, and Implementation*. New York: The Guilford Press.