EXPERIMENT 1: SCALE DEVELOPMENT

For each bias, we conducted a literature review to identify canonical questions, paradigms, and generated similar additional items (approximately 200 in total). BBS questions were developed following the question format of Scopelliti and colleagues (2015). FAE questions were based on the attitude attribution, quizmaster, silent interview, and moral attribution paradigms (Gawronski, 2004). CB questions were developed based on six paradigms: Wason's (1960) card selection task, Wason's (1968) triplets task; Tschirgi's (1980) cause identification paradigm, Snyder and Swann's (1978) trait hypothesis testing paradigm, an enriched versus impoverished profiles choice paradigm (Downs & Shafir, 1999), and a judgment of covariation paradigm (Nisbett & Ross, 1980). Three interchangeable versions (i.e., subscales) were created for each scale, so that each participant would see different questions at pretest (before training), posttest (immediately after training), and follow-up (8 weeks after training).

One sample of 288 Amazon Mechanical Turk (AMT) workers answered all FAE and BBS items. A separate sample of 310 AMT workers answered all CB items. We performed scale purification using an iterative procedure. In order to ensure that three valid and interchangeable versions of each scale were developed, questions with low item-total correlations were removed until random sampling suggested that a subsample of one third of the items on each bias scale would achieve a minimum of $\alpha \geq .7$ reliability at least 95% of the time. This purification resulted in a 27-item BBS scale, a 45-item FAE scale, two 9-item scales based on Wason (1960, 1968), a 12-item scale based on Tschirgi (1980), and three 18-item scales based on Snyder and Swann (1978), Downs and Shafir (1999), and Nisbett and Ross (1980). Exploratory factor analyses of the purified scales

indicated a unidimensional structure for each scale, with average variance explained = 36%. All items correlated positively with their respective factor, with an average minimum $r = .41$.

Seven to 11 days after completing the full scales, 305 participants completed purified versions of the same scales. Responses indicated high test-retest reliability and stability over time, $M_r = .79$. We divided each scale into three interchangeable subscales by iteratively selecting the three items with the highest average correlations and placing them into separate subscales, all subscale α's $\geq .65$. Items were divided among subscales so that subscales were maximally similar.

Item scoring logic varied due to their different formats. All item scores varied between 0 and 1, with 1 indicating greater bias (i.e., choosing confirming answers, making dispositional attributions, indicating less susceptibility to bias than one's peers). We calculated subscale scores by summing all individual items (i.e., all BBS items, FAE items, or CB items) and transforming totals into a score ranging from 0 (no biased answers) to 100 (all answers biased).

Bias knowledge questions had two forms. Recognition questions described an instance in which one of the three biases was committed and required participants to identify the bias in a free recall format. Discrimination questions described an instance of bias and tested its identification in a multiple-choice format. The final questionnaires contained 24 questions with satisfactory face validity (12 for recognition and 12 for discrimination), equally divided among the three biases.

EXPERIMENT 2: SCALE DEVELOPMENT

Three interchangeable subscales were created for each scale, so that each participant would see different questions at pretest (before training), posttest (immediately after training), and follow-up (8 weeks after training). For each bias, we conducted a literature review to identify canonical questions, paradigms, and generated similar additional items (423 in total). Anchoring questions used self-generated or experimenter provided anchors that were relevant or irrelevant (Strack & Mussweiler, 1997; Tversky & Kahneman, 1974; Simmons, Nelson, & LeBoeuf, 2010). Projection questions were developed from three bias facets: the false consensus effect (Ross, Greene, & House, 1976), attributive similarity (Holmes, 1968; Kreuger & Stanke, 2001), and the curse of knowledge (Birch & Bloom, 2007). The curse of knowledge dimension was not included in the final instrument based on factor analyses suggesting its exclusion. Representativeness questions were based on conjunction fallacy, base rate neglect, gambler's fallacy, perceptions of random sequences, and sample size neglect paradigms (Tversky & Kahneman, 1974).

After an initial purification stage, three samples of AMT workers ($N = 624$) completed the scales. Purification resulted in a 54-item anchoring scale, a 69-item projection scale, and a 78-item representativeness scale. Questions were then split into three interchangeable subscales for each bias. All the subset scales had acceptable internal consistency, $M_\alpha = .68$, and test re-test reliability, $M_r = .61$. Item and subscale scoring logic followed the procedure used in Experiment 1.

All bias knowledge questions for Experiment 2 were multiple choice discrimination questions. The final questionnaire contained 21 questions with satisfactory

face validity, divided into subscales with 7 questions each. Knowledge scales were scored on a 0-100 scale with higher scores indicating greater knowledge.

References

Birch, S. A., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science*, *18*, 382-386.

Gawronski, B. (2004). Theory-based bias correction in dispositional inference: The fundamental attribution error is dead, long live the correspondence bias. *European review of social psychology*, *15*(1), 183-217.

Downs, J. S., & Shafir, E. (1999). Why some are perceived as more confident and more insecure, more reckless and more cautious, more trusting and more suspicious, than others: Enriched and impoverished options in social judgment. *Psychonomic bulletin & review*, *6*, 598-610.

Holmes, D. S. (1968). Dimensions of projection. *Psychological Bulletin*, *69*(4), 248-268.

Krueger, J., & Stanke, D. (2001). The role of self-referent and other-referent knowledge in perceptions of group characteristics. *Personality and Social Psychology Bulletin*, *27*(7), 878-888.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice Hall.

Scopelliti, I., Morewedge, C. K., McCormick, E., Min, H. L., Lebrecht, S., & Kassam, K. S. (2015). Bias Blind Spot: Structure, Measurement, and Consequences. *Management Science*.

Simmons, J. P., LeBoeuf, R. A., & Nelson, L. D. (2010). The effect of accuracy motivation on anchoring and adjustment: do people adjust from provided anchors?. *Journal of Personality and Social Psychology*, *99*, 917-932.

Snyder, M., & Swann, W. B. (1978). Hypothesis-testing processes in social interaction. *Journal of Personality and Social Psychology*, *36*(11), 1202-1212.

Strack, F., & Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology*, *73*(3), 437-446.

Tschirgi, J. E. (1980). Sensible reasoning: A hypothesis about hypotheses. *Child development*, 1-10.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly journal of experimental psychology*, *12*(3), 129-140.

Wason, P. C. (1968). Reasoning about a rule. *The Quarterly journal of experimental psychology*, *20*(3), 273-281.