1 Summary

Supplementary materials for "Detecting social biases using mental state inference" contain additional details about the models, and methods and analyses for the experiments.

2 Computational Model

We use the same four models, one main ("Full") model and three alternative models, to generate predictions for all experiments. Here we provide specifications and implementation details for each model. Full model code is available at `https://osf.io/kprz3/?view_only=25d06dd71efb4b8699ab151ab40e4bc5`. All models were implemented in WebPPL, a probabilistic programming language for generative models.

2.1 Main model

A single forward pass of our main generative model comprises the following steps:

1. Sample the actor's true competence c_{true} from a symmetric Beta distribution $Beta(5,5)$.

2. Compute the actor's probability of success on each task (easy, medium, hard), based on the sample competence value c_{true} and the difficulty value of each task, which are fixed to $d_{easy} = .1$, $d_{medium} = .3$, and $d_{hard} = .7$. On a single attempt, the actor will get lucky and succeed with probability .1, get unlucky and fail with probability .1, or succeed with probability $1/(1+exp(\beta(c_{true}-d))$. This is a standard logit transformation with temperature parameter β, which we set equal to 4 based on maximum likelihood estimation from an earlier pilot study.

3. Sample the observer's prior belief distribution over the actor's competence. We define three possible prior distributions: $P_{pos} = Beta(4,1)$ is skewed towards higher values, $P_{neut} = Beta(5,5)$ is symmetric, and $P_{neg} = Beta(1,4)$ is skewed towards lower values. We sample the observer's prior $Prior_{obs}$ uniformly at random from these three priors.

4. Compute the observer's posterior belief distribution over the actor's competence given the observed performance O_{obs}. We do this via Bayesian updating, i.e.:

$$P(comp|Prior_{obs}, O_{obs}) \propto P(O_{obs}|comp)P(comp|Prior_{obs})$$

   The likelihood term $P(O_{obs}|comp)$ is given by the success probability computation in step 2.

5. Sample the observer's updated competence belief c_{obs} from the posterior distribution obtained in step 4.

6. Compute the expected point value of a single attempt on each task, given the coach's competence belief c_{obs}. For each task i, this is equal to the probability of success on task i (computed in step 2) times the point value of each task (1 for easy, 2 for medium, 3 for hard).

7. Given the observer's expected point values for each task (v_{easy}, v_{med}, v_{hard}), we compute the observer's probability of recommending each task by applying a SoftMax transformation with temperature parameter $\beta = 2$ (also tuned via Maximum Likelihood estimation from a previous pilot study) to the expected point values.

These steps define a joint probability distribution over the actor's true competence, the observer's prior belief distribution, and the observer's posterior competence belief, given observed values for the actor's performance and the observer's recommendation. For competence predictions, we output the expected value of the marginal distribution over the appropriate competence type (true or coach's posterior belief). For prior belief predictions, we output the inferred probability that the observer held each prior belief.

2.1.1 Implementation

To enable efficient, analytic computations of all posterior distributions, we approximated the prior distributions with discretized "Beta" distributions defined as follows: for each value x in the set $\{0, .01, .02, \ldots, .99, 1\}$, we computed the density of x under the corresponding Beta distribution, then normalized these values to obtain a discrete probability distribution over $\{0, .01, \ldots, 1\}$ that approximated the corresponding Beta distribution. This enabled us to compute the posterior distributions analytically, rather than having to approximate via sampling.

2.2 Alternate models

2.2.1 Recommendation only model

The *Recommendation only model* omits steps 4 and 5 from the main model: rather than updating the observer's competence belief on the observed performance O_{obs}, we sample the observer's posterior competence belief directly from the prior. The remaining model is left unchanged. This simulates inference under the assumption that the observer's recommendation is not influenced by the observed performance.

2.2.2 Discounting model

The *Discounting model* follows the same steps as the full model, but modifies the posterior belief update process in Step 4. First, for each observation, it estimates $P(luck = 1 | difficulty, outcome, prior)$, i.e.: the probability that the actor got lucky on that attempt, based on the observer's prior belief distribution (sampled in step 3). We do this by sampling 10,000 actor competence values from the observer's prior, simulating an attempt from each actor on the appropriate task, the computing the fraction of samples in which a) the attempt yielded the outcome seen in the observation and b) the actor got lucky. The model then either ignores that observation (i.e.: does not update the posterior) with probability $P(luck = 1 | difficulty, outcome, prior)$, or updates the posterior on that observation with probability $1 - P(luck = 1 | difficulty, outcome, prior)$. This simulates inference

under the assumption that the observer is more likely to ignore evidence that reflects luck rather than skill (and that a more biased observer is more likely to dismiss successes as being luck-based). The remaining model is left unchanged.

2.2.3 Observation only model

The *Observation only model* omits steps 6 and 7 from the main model, which yields a joint distribution over true competence, the observer's prior belief, and the observer's posterior belief (but not the observer's recommendation). By computing the competence belief and prior belief from this truncated distribution, the model simulates inference under the assumption that the observed performance directly influences the observer's prior bias (e.g.: that observing a bad performance causes a negative bias).

2.3 Pre-registered versus published model

The model that we pre-registered is slightly different than the model we used to generate the predictions reported in the main manuscript. The reason is as follows: after pre-registering the initial model predictions and running our pre-registered experiments, we discovered a subtle bug in the code of our original model. This bug created a statistical dependency between two variables that should not have been there. In particular, under the pre-registered model, the observer's prior belief was directly statistically dependent on the actor's performance. That is, the model predicted that the observer's pre-existing bias was retroactively influenced by observing the actor's performance. According to the causal model we proposed, however (see Figure 1 in the main manuscript), the observer's prior belief should only be dependent on the actor's performance when conditioning on the observer's recommendation. Thus, due to this bug, the model we originally pre-registered did not faithfully implement the causal model that we hypothesized.

After discovering and correcting this bug, we re-tuned our two free parameters against the original pilot data and re-generated all model predictions with these new parameters (actor softMax value: 4, observer softMax value: 2). These predictions are the ones reported in our main

manuscript. Fortunately, this change had little impact on the overall model fit, leading to slight but not statistically significant improvements in model correlation for each experiment.

3 Experiment 1

3.1 Supplemental Methods

3.1.1 Comprehension check questions

After completing the 10 test trials, participants were asked four comprehension check questions to test their understanding of the cover story. Our preregistered exclusion criteria for all experiments stated that if participants responded incorrectly to one or more of these questions, they would be excluded from analyses. Note that participants underwent similar questions *before* they underwent the test trials, except the wording was in present tense rather than past tense as below. There, they were forced to get the questions correct and they had unlimited chances to do so. Below is the exact wording of the comprehension check questions used for exclusions, with the correct answer bolded.

1. At the beginning of the day, when each player said the name of their team, the coach:
    - Never heard it correctly
    - **Sometimes heard it correctly**
    - Always heard it correctly

2. When the player attempted a shot, was there a small chance that the player will get lucky or unlucky?
    - **Yes**
    - No

3. When the coach suggested a basket for the player to do, what was the coach's goal?
    - **For the player to get the most points**
    - For the player to get better

- For the player to work hard
- For the player to feel good about themselves

4. When the coach suggested a basket for the player to do, what did the coach consider?
    - The team that the coach heard
    - Their observations of the player
    - **Both the team that the coach heard and their observations**

3.2 Supplementary Results

3.2.1 Preregistered model results

Our preregistered Full model showed an overall strong quantitative fit with participant judgments (r=.90, 95% CI: [.86, .92]), and model fit was similar for each inference type (competence: r=.95, 95% CI [.92, .97]; team: r=.87, 95% CI [.82, .92]). The *Recommendation only model* had an overall correlation of r=.90 (95% CI: [.86,.92]), which was not reliably different from the Full model (Δ = .00, 95% CI: [-.03, .02]). The *Observation only model* had an overall correlation of r=.60 (95% CI: [.49,.69]), which was reliably lower than the Full model (Δ = .30, 95% CI: [.20, .42]).

4 Experiment 2

4.1 Supplemental Methods

4.1.1 Comprehension check questions

Similar comprehension check questions were used for Experiment 2 as in Experiment 1, with minor changes given the cover story.

1. At the beginning of the day, when the coaches met the players, the coaches:
    - Always formed negative biases
    - **Formed negative or positive biases, or no biases**

- Always formed positive biases

2. When the player attempted a shot, was there a small chance that the player will get lucky or unlucky?

- **Yes**
- No

3. When the coach suggested a basket for the player to do, what was the coach's goal?

- **For the player to get the most points**
- For the player to get better
- For the player to work hard
- For the player to feel good about themselves

4. When the coach suggested a basket for the player to do, what did the coach consider?

- Their initial bias
- Their observations of the player
- **Both their initial bias and their observations**

4.2 Supplementary Results

4.2.1 Preregistered model results

Our preregistered *Full model* showed a strong quantitative fit with participant judgments (r=.79, 95% CI: [.72, 84]). Model fit was high for each inference type (competence: r=.90, 95% CI: [.84, 94]); bias: r=.74, 95% CI: [.63, 82]).

The *Recommendation only model* had an overall correlation of r=.73, 95% CI: [.64, .80], which was reliably lower than the Full model (Δ=.05, 95% CI: [.02, .09]). The *Observation only model* had an overall correlation of r=.78, 95% CI: [.71, .83], which was not reliably lower than the Full model (Δ=.008, 95% CI: [-.05,.07].

5 Experiment 3

5.1 Supplemental Methods

5.1.1 Comprehension check questions

Similar comprehension check questions were used for Experiments 3a-3c as in Experiment 2, with minor changes given the cover story.

Experiment 3a:

1. At the beginning of the day, when the coaches met the players, the coaches:
    - Always formed negative biases
    - **Formed negative or positive biases, or no biases**
    - Always formed positive biases
2. When the coach suggested a basket for the player to do, what was the coach's goal?
    - **For the player to get the most points**
    - For the player to get better
    - For the player to work hard
    - For the player to feel good about themselves
3. When the coach suggested a basket for the player to do, what did the coach consider?
    - Their initial bias
    - Their observations of the player
    - **Both their initial bias and their observations**

Experiment 3b:

1. At the beginning of the day, when the teachers met the students, the teachers:
    - Always formed negative biases
    - **Formed negative or positive biases, or no biases**
    - Always formed positive biases
2. When the teacher suggested an extra credit problem for the student to do, what was the teacher's goal?

- **For the student to get the most points**
- For the student to get better
- For the student to work hard
- For the student to feel good about themselves

3. When the teacher suggested a problem for the student to do, what did the teacher consider?
    - Their initial bias
    - Their observations of the student
    - **Both their initial bias and their observations**

Experiment 3c:

1. At the beginning of the day, when the recruiters met the potential hires, the recruiters:
    - Always formed negative biases
    - **Formed negative or positive biases, or no biases**
    - Always formed positive biases
2. When the recruiter suggested a problem the candidate to do, what was the recruiter's goal?
    - **For the candidate to get the most money**
    - For the candidate to get better
    - For the candidate to work hard
    - For the candidate to feel good about themselves
3. When the recruiter suggested a problem for the student to do, what did the recruiter consider?
    - Their initial bias
    - Their observations of the candidate
    - **Both their initial bias and their observations**

5.1.2 Stereotype endorsement questions

At the end of each experiment, participants were asked about their agreement with the stereotypes relevant to the cover story. Below is the exact wording and response options for this question:

Experiment 3a: There are stereotypes about the abilities of different social groups (e.g., gen-

der, race, class, age). Some stereotypes are accurate and some are inaccurate. Think about this stereotype: Black people are more skilled at basketball than White people. How accurate is this stereotype?

- Not at all accurate
- Slightly accurate
- Moderately accurate
- Very accurate
- Extremely accurate

Experiment 3b: There are stereotypes about the abilities of different social groups (e.g., gender, race, class, age). Some stereotypes are accurate and some are inaccurate. Think about this stereotype: Boys are more skilled at math than girls. How accurate is this stereotype? (same response options as above)

Experiment 3c: There are stereotypes about the abilities of different social groups (e.g., gender, race, class, age). Some stereotypes are accurate and some are inaccurate. Think about this stereotype: Young people are more skilled at programming than old people. How accurate is this stereotype? (same response options as above)

5.2 Supplemental Results

5.2.1 Models presented in paper

For each experiment, the *Full model* produced high fits for the competence inferences (3a: r=.95, 95% CI [.89, .97]; 3b: r=.94, 95% CI [.87, .97]; 3c: r=.81, 95% CI [.63, .91]) and the bias inferences (3a: r=.76, 95% CI [.66, .84]; 3b: r=.86, 95% CI [.80, .91]; 3c: r=.73, 95% CI [.62, .82]).

The *Full model* produced reliably higher fits compared to the *Discounting model* (Experiment 3a: $|\Delta|$ = .14, 95% CI [.10, .19]; Experiment 3b: $|\Delta|$ = .07, 95% CI [.05, .11]; Experiment 3c: $|\Delta|$ = .14, 95% CI [.10, .19]), *Recommendation only model* (Experiment 3a: $|\Delta|$ = .02, 95% CI [.01, .04]; Experiment 3b: $|\Delta|$ = .01, 95% CI [.00, .02]; Experiment 3c: $|\Delta|$ = .03, 95% CI [.02, .05]),

and the *Observation only model* (Experiment 3a: $|\Delta|$ = .34, 95% CI [.25, .45]; Experiment 3b: $|\Delta|$ = .50, 95% CI [.37, .65]; Experiment 3c: $|\Delta|$ = .25, 95% CI [.18, .33]).

Further, to test whether the Full model captures variability that the *Observation only model* does not, we ran correlations between the Full model and average participants' responses for sets of trials for which the *Observation only model* makes identical predictions (24 sets of trials x 3 experiments = 72 correlations). We found that 72 of the 72 sets of trials produced positive correlations (100% of trials; $p < .001$, Binomial Test). Thus, the Full model captured meaningful variability that the *Recommendation only model* (see results reported in main paper) and *Observation only model* did not.

5.2.2 Preregistered model results

The Full model showed a strong quantitative fit with participant judgments for Experiment 3a (r=.81, 95% CI: [.74, .86]), Experiment 3b (r=.83, 95% CI: [.77, .87]), and Experiment 3c (r=.83, 95% CI: [.77, .87]; preregistered). Across experiments, model fit was high for the inferences about the observer's beliefs about the actor (Experiment 3a: r=.94, 95% CI: [.87, .97]; Experiment 3b: r=.95, 95% CI: [.91, .98]; Experiment 3c: r=.93, 95% CI: [.85, .97]) and inferences about the observer's biases (Experiment 3a: r=.76, 95% CI: [.66, .84]; Experiment 3b: r=.86, 95% CI: [.80, .91]; Experiment 3c: r=.73, 95% CI: [.62, .82]).

For all experiments, the Full model produced reliably higher fits compared to the Observation model (Experiment 3a: Δ=.12, 95% CI: [.04, .22]; Experiment 3b: Δ=.29, 95% CI: [.17, .41]; Experiment 3c: Δ=.09, 95% CI: [.03, .17]).

However, the Full model did not consistently do better than the Recommendation model (Experiment 3a: Δ=.004, 95% CI: [-.03, .03]; Experiment 3b: Δ=-.04, 95% CI: [-.07, -.01]; Experiment 3c: Δ=.04, 95% CI: [.01, .07]). To test whether the Full model captures variability that the *Recommendation only model* does not, we ran correlations between the Full model and average participants' responses for sets of trials for which the *Recommendation only model* makes identical predictions (12 sets of trials x 3 experiments = 36 correlations; preregistered analysis). We

found a positive correlation for 28 of the 36 sets of trials (77.78% of trials; $p = .001$, Binomial Test). We preregistered running the same analysis comparing the Full Model and the Observation model (correlations between the Full model and average participants' responses for sets of trials for which the *Observation only model* makes identical predictions; 24 sets of trials x 3 experiments = 72 correlations). We found that 71 of the 72 sets of trials produced positive correlations (98.61% of trials; $p < .001$, Binomial Test). Thus, the Full model captured meaningful variability that the *Recommendation only model* and *Observation only model* did not.

6 Experiment 4

6.1 Supplemental Methods

6.1.1 Comprehension check questions

Similar comprehension check questions were used for Experiment 2 as in Experiment 1, with minor changes given the cover story.

1. When the student answers a question, is there a small chance that the student got lucky or unlucky?

- **Yes**
- No

2. When the counselor recommends a course, what is their goal?

- **For the student to take the hardest course that they can succeed in**
- For the student to take a really fun class
- For the student to take a course that they can improve in

3. The counselor also learns the gender and race of the student. The counselor:

- Never holds stereotypes about the student's gender or racial group
- **Sometimes holds stereotypes about the student's gender or racial group**
- Always holds stereotypes about the student's gender or racial group

4. What might the counselor consider when making their recommendation?

- The student's race/gender
- The student's scores
- Both the student's race/gender and their scores

7 Tables

Trial	Easy task	Medium task	Hard task	Recommendation
1	**Miss**, Hit	Miss	Miss	[Easy, Medium, or Hard]
2	Miss, **Hit**	Miss	Miss	[Easy, Medium, or Hard]
3	Hit	**Miss**, Hit	Miss	[Easy, Medium, or Hard]
4	Hit	Miss, **Hit**	Miss	[Easy, Medium, or Hard]
5	Hit	Hit	**Miss**, Hit	[Easy, Medium, or Hard]
6	Hit	Hit	Miss, **Hit**	[Easy, Medium, or Hard]
7	**Miss**	Hit	Hit, Hit	[Easy, Medium, or Hard]
8	Miss	Hit	**Hit**, Hit	[Easy, Medium, or Hard]
9	**Miss**, Miss	Miss	Hit	[Easy, Medium, or Hard]
10	Miss, Miss	Miss	**Hit**	[Easy, Medium, or Hard]

Table S1: Trials for Experiments 1-3. Each trial showed the actor attempting four shots, and the observer saw one of the shots (bolded). Then, the observer suggested the easy, medium, or hard shot. There were 10 different patterns of the actor's performance and the observer's observations; for each of these, we created three trials that varied the observer's recommendation, for 30 trials total. We created 3 sets of 10 trials, where each set contained all 10 variations of actor performance and observer observation and varied the observer recommendation. Each participant underwent one of these sets (10 trials; randomized order).

Trial	Demographics	Math Results	Nursing Results	Math Rec.	Nursing Rec.
1	Asian Female	F, F, S	F, F, F	Advanced	Advanced
2	Asian Male	F, F, F	F, F, F	Advanced	Beginner
3	Latine Female	F, S, F	F, S, F	Beginner	Advanced
4	Black Male	F, S, S	F, S, S	Advanced	Beginner
5	Asian Male	F, S, S	F, S, S	Advanced	Beginner
6	Black Female	F, F, F	F, F, F	Beginner	Advanced
7	Asian Female	F, F, F	F, F, S	Advanced	Advanced
8	Asian Male	F, F, F	S, S, F	Beginner	Beginner
9	Latine Female	S, S, S	F, S, F	Intermediate	Intermediate
10	Black Male	F, F, F	S, S, S	Beginner	Beginner
11	Latine Female	S, S, S	F, F, F	Advanced	Advanced
12	Asian Female	F, F, F	F, S, F	Advanced	Advanced
13	Asian Male	F, F, S	S, S, S	Beginner	Beginner
14	Black Female	F, F, S	F, S, S	Beginner	Beginner
15	Latine Female	F, S, F	F, S, F	Beginner	Intermediate

Table S2: Trials for Experiment 4. Each trial showed a student's race/ethnicity and gender information, and their performance on three math questions and three nursing questions. Participants either failed (F) or succeeded (S) on each question, and were shown a beginner, intermediate, and advanced question for both math and nursing. For example, in the above table, "F, S, F" means that the student failed a beginner question, succeeded on an intermediate question, and a failed an advanced question. The college counselor observed all of the student's performance and then recommended that the student enroll in a beginner, intermediate, or advanced math course and a beginner, intermediate, or advanced nursing course. We created 3 sets of 5 trials; the first set comprised Trials 1-5, the second set comprised Trials 6-10, and the third set comprised Trials 11-15. Each participant was randomly assigned to one of the sets and saw 5 trials in a randomized order.