Awareness of implicit attitudes:

Large-scale investigations of mechanism and scope

Online Supplement

Adam Morris[1*] and Benedek Kurdi[2]

[1] Department of Psychology, Princeton University

[2] Department of Psychology, Yale University

[*] To whom correspondence should be addressed: adam.mtc.morris@gmail.com

Differences Between Study 1B and the Paradigm by Hahn et al. (2014)

In spite of a high level of procedural similarity, Study 1B of the present project produced a considerably smaller effect than the original studies by Hahn et al. (2014). The most obvious difference between the two sets of studies is that the present work was conducted online in a relatively diverse sample of adult participants, whereas the work by Hahn et al. (2014) was conducted in person in a relatively homogeneous sample of student participants.

Nonetheless, there are other minor procedural differences between the two studies some of which may have contributed to the difference in effect size:

(1) Unlike Hahn et al. (2014) who used full IATs, the present work relied on an abbreviated IAT design without practice blocks. It should be noted, however, that mean level IAT D scores were highly correlated across the present studies and studies by Nosek (2005; see Studies 2A–2C). As such, we believe that this aspect of the design is unlikely to be responsible for the difference.

(2) Within the prediction items, images representing the two categories were presented above and below each other instead of grouped to the left and right, as they had been in work by Hahn et al. (2014).

(3) There were other minor visual differences between the two studies in terms of font size, font color, etc.

(4) In the present studies, predictions were made on 100-point sliding scales instead of 7-point Likert scales.

(5) Explicit attitudes were assessed post-prediction. This difference was empirically addressed in Studies 2B–2C and 3B and was found not to account for the observed difference.

Final Model Specifications

Here we include model specifications (including random effects structures) for each final (best-fitting) model reported in the paper. The dependent variable in each case was the IAT D score, with the exception of Studies 3A–3B where the dependent variable was the AMP score. None of the models included participant-level random intercepts because all variables were standardized within participants.

Study 1A

Model 1. This model included a fixed effect for the prediction item, as well as prediction-by-participant random slopes, random intercepts for targets, and prediction-by-target random slopes:

```
dScoreS ~ predictS + (0 + predictS | session_id) + (predictS | comp)
```

Model 2. This model included fixed effects for the prediction item and explicit attitudes, as well as prediction-by-participant random slopes, random intercepts for targets, and prediction-by-target random slopes:

```
dScoreS ~ predictS + expDiffS + (0 + predictS | session_id) + (predictS | comp)
```

Study 1B

Model 1. This model included a fixed effect for the prediction item, as well as prediction-by-participant random slopes and random intercepts for targets:

```
dScoreS ~ predictS + (0 + predictS | session_id) + (1 | comp)
```

Model 2. This model included fixed effects for the prediction item and explicit attitudes, as well as prediction-by-participant random slopes and random intercepts for targets:

```
dScoreS ~ predictS + expDiffS + (0 + predictS | session_id) + (1 | comp)
```

Study 2A

Model 1. This model included a fixed effect for the prediction item, as well as prediction-by-participant random slopes and prediction-by-target random slopes:

```
dScoreS ~ predictS + (0 + predictS | session_id) + (predictS | comp)
```

Model 2. This model included fixed effects for the prediction item and explicit attitudes, as well as prediction-by-participant random slopes, random intercepts for targets, and (uncorrelated) prediction-by-target random slopes:

```
dScoreS ~ predictS + expDiffS + (0 + predictS | session_id) + (predictS || comp)
```

Study 2B

Model 1. This model included a fixed effect for the prediction item, as well as prediction-by-participant random slopes and prediction-by-target random slopes:

```
dScoreS ~ predictS + (0 + predictS | session_id) + (predictS | comp)
```

Model 2A. This model included fixed effects for the prediction item and time-2 explicit attitudes, as well as prediction-by-participant random slopes, random intercepts for targets, and prediction-by-target random slopes:

```
dScoreS ~ predictS + expDiffT2S + (0 + predictS | session_id) + (predictS | comp)
```

Model 2B. This model included fixed effects for the prediction item and time-1 explicit attitudes, as well as prediction-by-participant random slopes and prediction-by-target random slopes:

```
dScoreS ~ predictS + expDiffT1S + (0 + predictS | session_id) + (predictS | comp)
```

Model 2C. This model included fixed effects for the prediction item and time-1 and time-2 explicit attitudes, as well as prediction-by-participant random slopes and prediction-by-target random slopes:

```
dScoreS ~ predictS + expDiffT1S + expDiffT2S + (0 + predictS | session_id) + (predictS | comp)
```

Study 2C

Model 1. This model included a fixed effect for the prediction item, as well as prediction-by-participant random slopes, random intercepts for targets, and (uncorrelated) prediction-by-target random slopes:

```
dScoreS ~ predictS + (0 + predictS | session_id) + (predictS || comp)
```

Model 2A. This model included fixed effects for the prediction item and time-2 explicit attitudes, as well as prediction-by-participant random slopes, random intercepts for targets, and (uncorrelated) prediction-by-target random slopes:

```
dScoreS ~ predictS + expDiffT2S + (0 + predictS | session_id) + (predictS || comp)
```

Model 2B. This model included fixed effects for the prediction item and time-1 explicit attitudes, as well as prediction-by-participant random slopes, random intercepts for targets, and prediction-by-target random slopes:

```
dScoreS ~ predictS + expDiffT1S + (0 + predictS | session_id) + (predictS | comp)
```

Model 2C. This model included fixed effects for the prediction item and time-1 and time-2 explicit attitudes, as well as prediction-by-participant random slopes, random intercepts for targets, and (uncorrelated) prediction-by-target random slopes:

```
dScoreS ~ predictS + expDiffT1S + expDiffT2S + (0 + predictS | session_id) + (predictS || comp)
```

Models 3A–3C. These models had the same random effects structure as the underlying models 2A–2C but included interactions of each predictor with type of explicit item:

```
dScoreS ~ predictS * expType + expDiffT2S * expType + (0 +
predictS | session_id) + (predictS || comp)
dScoreS ~ predictS * expType + expDiffT1S * expType + (0 +
predictS | session_id) + (predictS | comp)
dScoreS ~ predictS * expType + expDiffT1S * expType +
expDiffT2S * expType + (0 + predictS | session_id) + (predictS
|| comp)
```

Study 3A

Model 1. This model included a fixed effect for the prediction item, as well as prediction-by-participant random slopes and random intercepts for targets:

```
ampScoreS ~ predictS + (0 + predictS | session_id) + (1 |
comp)
```

Model 2. This model included fixed effects for the prediction item and explicit attitudes, as well as prediction-by-participant random slopes and random intercepts for targets:

```
ampScoreS ~ predictS + expDiffS + (0 + predictS | session_id)
+ (1 | comp)
```

Study 3B

Model 1. This model included a fixed effect for the prediction item, as well as prediction-by-participant random slopes, random intercepts for targets, and prediction-by-target random slopes:

```
ampScoreS ~ predictS + (0 + predictS | session_id) + (predictS
| comp)
```

Model 2A. This model included fixed effects for the prediction item and time-2 explicit attitudes, as well as prediction-by-participant random slopes, random intercepts for targets, and prediction-by-target random slopes:

```
ampScoreS ~ predictS + expDiffT2S + (0 + predictS |
session_id) + (predictS | comp)
```

Model 2B. This model included fixed effects for the prediction item and time-1 explicit attitudes, as well as prediction-by-participant random slopes, random intercepts for targets, and prediction-by-target random slopes:

```
ampScoreS ~ predictS + expDiffT1S + (0 + predictS |
session_id) + (predictS | comp)
```

Model 2C. This model included fixed effects for the prediction item and time-1 and time-2 explicit attitudes, as well as prediction-by-participant random slopes, random intercepts for targets, and prediction-by-target random slopes:

```
ampScoreS ~ predictS + expDiffT1S + expDiffT2S + (0 + predictS
| session_id) + (predictS | comp)
```

Study 4

Model 1. This model included a fixed effect for the third-party prediction item, as well as prediction-by-participant random slopes for both observers and deciders, random intercepts for attitude targets, and prediction-by-target random slopes:

```
dScoreOrigS ~ predictS + (0 + predictS | observer_session_id)
+ (0 + predictS | decider_session_id) + (predictS | comp)
```

Model 2. This model included fixed effects for the third-party and first-person prediction items, as well as third-party prediction and first-person prediction-by-target participant random slopes (for deciders; including observer-level random slopes prevented convergence), and third-party prediction and first-person prediction-by-target random slopes:

```
dScoreOrigS ~ predictS + predictOrigS + (0 + predictS +
predictOrigS | decider_session_id) + (predictS + predictOrigS
| comp)
```

Aggregate analysis

This model regressed predictions on implicit attitude scores, explicit attitudes, and demographic-based predictions; it included fixed effects for each of those three variables, subject-level random slopes for the first two, and item-level random slopes for all three plus an item-level random intercept:

```
predictS ~ dScoreS + expDiffS + demopred + (0 + dScoreS +
expDiffS | session_id) + (dScoreS + expDiffS + demopred |
comp)
```

Moderators from Nosek (2005)

Nosek (2005) measured, for each attitude target, four key variables that could characterize differences between attitudes: how much the attitude elicits self-presentation concerns ("self-presentation"); how practiced people are at evaluating the target ("evaluative strength"); how simple or complex the representation of the target's valence is ("dimensionality"); and how much people perceive their evaluations of the target to diverge from normative evaluations ("distinctiveness"). In Studies 2A–2C and 3A–3B, we tested whether these variables accounted for any variance in target-level variation in predictive accuracy.[1]

Study 2A

When modeling the zero-order relationship between predictions and IAT scores as the dependent variable, we obtained a large and significant effect for self-presentation, β = -0.781,

[1] It should be noted, however, that Nosek (2005) measured these properties at the participant level. For our purposes, we focus on the average value of each property at the attitude target level. Future work could investigate the relationship between these variables and predictive accuracy at the participant level, which may well be different from the target-level findings reported here.

$t(49) = -5.29$, $p < .001$, and a medium-sized and significant effect for evaluative strength, $\beta = 0.339$, $t(49) = 2.12$, $p = .039$. That is, predictions were more likely to be accurate for less socially sensitive topics and for topics subject to more evaluative elaboration. The remaining two variables did not produce significant effects.

However, when analyzing the uniquely implicit component of predictive accuracy, none of the four variables were significantly predictive. Moreover, the overall model fit the data considerably worse: There was a drop from an R^2 of 0.50 when modeling the zero-order relationship to an R^2 of 0.16 when modeling the uniquely implicit component of the relationship.

Study 2B

When modeling the zero-order relationship between predictions and IAT scores as the dependent variable, we obtained a mid-sized and significant effect for self-presentation, $\beta = -0.446$, $t(47) = -2.53$, $p = 0.015$. The remaining three variables (including evaluative strength, which was significant in Study 2A) did not produce significant effects.

Similar to Study 2A, when analyzing the uniquely implicit component of predictive accuracy, none of the four variables were significantly predictive. Moreover, the overall model fit the data considerably worse: There was a drop from an R^2 of 0.32 when modeling the zero-order relationship to an R^2 of 0.10 (time-1 explicit attitudes) and 0.03 (time-2 explicit attitudes) when modeling the uniquely implicit component of the relationship.

Study 2C

No effects were significant in any of the three models.

Study 3A

No effects were significant in either model.

Study 3B

No effects were significant in any of the three models.

Discussion

The results obtained in Study 2A suggest that self-presentation concerns and evaluative strength are associated with people's ability to predict the portion of their implicit attitudes that overlap with their explicit attitudes, but that these variables are unrelated to people's ability to predict the uniquely implicit component. Whereas Study 2B replicated this result with respect to self-presentation, the effect of evaluative strength did not replicate. In the remaining studies, no effects were significant. However, at least in Studies 3A–3B, statistical power was poor as a result of the relatively small number of attitude targets included.

To summarize, these results potentially offer some insight into attitude target-level variability in predictive accuracy. However, the findings seem less robust across studies than would be desirable to be able to draw firm conclusions from them. Moreover, in any case, none of the moderator variables drawn from Nosek (2005) explained significant amounts of variance in the uniquely implicit component of predictive accuracy (after accounting for explicit attitudes).

Mixed-Effects Models with No Item-Level Random Effects

In line with the relevant preregistration, our analytic strategy for the models reported in the main text was to fit mixed-effects models with the maximal random effects structure and then, if necessary, simplify that random-effects structure on the basis of a principal component analysis of random effects. Specifically, random effects were removed if they accounted for 0% of the variance after rounding to three decimal places.

This analytic strategy deviates from the analytic strategy pursued by Hahn et al. (2014) who did not include any target-level random effects in their mixed effects models. Below we explore whether this difference in analytic strategies can account for the differences observed in the results of Hahn et al. (2014) and the present studies by refitting all relevant models from Studies 1–3 containing only prediction-by-participant random slopes but no random effects involving attitude targets.

Study 1A

In model 1, which used the prediction items as the only independent variable, we obtained a small but statistically significant relationship between the prediction items and IAT D scores, $\beta = 0.096$, $t(476) = 3.94$, $p < .001$. This result is in line with that reported in the main text.

In model 2, we found that the relationship between the prediction item and implicit attitudes persisted following the addition of the explicit attitude difference scores to the model. The effect size associated with the prediction item remained virtually unchanged, $\beta = 0.086$, $t(630.23) = 3.18$, $p = .002$. Unlike in the model reported in the main text, explicit attitudes did not produce a significant effect, $\beta = 0.021$, $t(2311.42) = 0.87$, $p = .386$.

Study 1B

In model 1, which used the prediction items as the only independent variable, we obtained a small but statistically significant relationship between the prediction items and IAT D scores, $\beta = 0.249$, $t(471) = 10.29$, $p < .001$. This result is qualitatively in line with that reported in the main text (although the predictive accuracy is larger).

In model 2, we found that the relationship between the prediction item and implicit attitudes persisted following the addition of the explicit attitude difference scores to the model. The effect size associated with the prediction item remained virtually unchanged, $\beta = 0.254$, $t(584.45) = 9.70$, $p < .001$. Unlike in the model reported in the main text, explicit attitudes did not produce a significant effect, $\beta = -0.002$, $t(2298.20) = -0.08$, $p = .938$.

Study 2A

In model 1, which used the prediction items as the only independent variable, we obtained a statistically significant and medium-sized relationship between the prediction items and IAT D scores, $\beta = 0.369$, $t(1837) = 33.87$, $p < .001$. This result is in line with that reported in the main text.

In model 2, we found that the relationship between the prediction item and implicit attitudes persisted following the addition of the explicit attitude difference scores to the model. Specifically, the effect of the prediction item remained significant, $\beta = 0.248$, $t(3662.14) = 18.00$, $p < .001$, and explicit attitudes also had a significant effect of a similar size, $\beta = 0.181$, $t(9095.32) = 13.98$, $p < .001$. These results are also in line with those reported in the main text.

Study 2B

In Study 2B, all models produced effects that were in line with the results reported in the main text.

In model 1, we obtained a statistically significant and medium-sized relationship between participants' predictions and IAT scores, $\beta = 0.337$, $t(1826) = 30.40$, $p < .001$.

In model 2A, we found that the relationship between the prediction item and implicit attitudes persisted following the addition of the time-2 explicit attitude difference scores to the model. Specifically, the prediction item had a reduced but significant effect, $\beta = 0.194$, $t(3766.80) = 13.59$, $p < .001$, and explicit attitudes also had a significant effect of similar size, $\beta = 0.210$, $t(9022.48) = 15.63$, $p < .001$.

In model 2B, a similar finding emerged when the time-1 instead of the time-2 explicit attitudes were used as controls: The prediction item had a reduced but significant effect, $\beta = 0.272$, $t(2921.74) = 21.06$, $p < .001$, and explicit attitudes also had a significant effect (although this time the effect was smaller), $\beta = 0.116$, $t(9014.45) = 9.66$, $p < .001$.

In model 2C, both sets of explicit items were considered simultaneously. In this model, the effect of the prediction item remained significant, $\beta = 0.194$, $t(3812.62) = 13.45$, $p < .001$, and time-2 explicit attitudes also had a significant effect, $\beta = 0.198$, $t(8950.92) = 12.09$, $p < .001$. The effect of time-1 explicit attitudes was not significant, $\beta = 0.014$, $t(8950.42) = 1.05$, $p = .296$.

Study 2C

Similar to Studies 2A–2B, all models fit to the data from Study 2C produced effects that were in line with the results reported in the main text.

In model 1, we obtained a statistically significant and medium-sized relationship between participants' predictions and IAT scores, $\beta = 0.342$, $t(1359) = 26.95$, $p < .001$.

In model 2A, we found that the relationship between the prediction item and implicit attitudes persisted following the addition of the time-2 explicit attitude difference scores to the model. Specifically, the prediction item had a reduced but significant effect, $\beta = 0.235$, $t(2434.45) = 15.18$, $p < .001$, and explicit attitudes also had a significant effect of similar size, $\beta = 0.174$, $t(6649.92) = 11.83$, $p < .001$.

In model 2B, a similar finding emerged when the time-1 instead of the time-2 explicit attitudes were used as controls: The prediction item had a reduced but significant effect, $\beta = 0.296$, $t(1937.00) = 20.68$, $p < .001$, and explicit attitudes also had a significant effect (although this time the effect was smaller), $\beta = 0.091$, $t(6685.28) = 6.82$, $p < .001$.

In model 2C, both sets of explicit items were considered simultaneously. In this model, the effect of the prediction item remained significant, $\beta = 0.232$, $t(2486.28) = 14.73$, $p < .001$, and time-2 explicit attitudes also had a significant effect, $\beta = 0.167$, $t(6626.08) = 9.64$, $p < .001$. The effect of time-1 explicit attitudes was not significant, $\beta = 0.014$, $t(6639.54) = 0.88$, $p = .377$.

Study 3A

Similar to Studies 2A–2C, both models fit to the data from Study 3A produced effects that were in line with the results reported in the main text.

In model 1, which used the prediction items as the only independent variable, we obtained a statistically significant and medium-sized relationship between the prediction items and AMP scores, $\beta = 0.435$, $t(358) = 16.71$, $p < .001$.

In model 2, we found that the relationship between the prediction item and implicit attitudes persisted following the addition of the explicit attitude difference scores to the model. Specifically, the effect of the prediction item remained significant, $\beta = 0.262$, $t(659.47) = 8.30$, $p < .001$, and explicit attitudes also had a significant effect of a similar size, $\beta = 0.261$, $t(1774.67) = 9.38$, $p < .001$.

Study 3B

Similar to Studies 2A–3A, all models fit to the data from Study 3B produced effects that were in line with the results reported in the main text.

In model 1, we obtained a statistically significant and medium-sized relationship between participants' predictions and AMP scores, $\beta = 0.416$, $t(529.81) = 19.55$, $p < .001$.

In model 2A, we found that the relationship between the prediction item and implicit attitudes persisted following the addition of the time-2 explicit attitude difference scores to the model. Specifically, the prediction item had a reduced but significant effect, $\beta = 0.266$, $t(917.00) = 10.62$, $p < .001$, and explicit attitudes also had a significant effect of similar size, $\beta = 0.241$, $t(2611.93) = 10.71$, $p < .001$.

In model 2B, a similar finding emerged when the time-1 instead of the time-2 explicit attitudes were used as controls: The prediction item had a reduced but significant effect, $\beta = 0.343$, $t(768.46) = 14.56$, $p < .001$, and explicit attitudes also had a significant effect (although this time the effect was smaller), $\beta = 0.139$, $t(2622.09) = 6.70$, $p < .001$.

In model 2C, both sets of explicit items were considered simultaneously. In this model, the effect of the prediction item remained significant, $\beta = 0.262$, $t(927.36) = 10.34$, $p < .001$, and time-2 explicit attitudes also had a significant effect, $\beta = 0.239$, $t(2591.25) = 8.40$, $p < .001$. The effect of time-1 explicit attitudes was not significant, $\beta = 0.010$, $t(2572.90) = 0.38$, $p = .701$.

Discussion

The omission of attitude target-level random effects did not change the results of Studies 2A–3C, which included a relatively large number of attitude objects. In Studies 1A–1B, unlike in the models reported in the main text, the incremental effect of explicit attitudes was not significant.

"Between-Subjects" Analyses

Following Hahn et al. (2014), our primary analyses tested the extent to which participants could predict the within-subject patterns of their implicit attitude scores by regressing participant-standardized implicit attitude scores on participant-standardized predictions. Here, again following Hahn et al. (2014), we additionally tested the extent to which participants could predict their absolute, "between-subjects" implicit attitude scores by regressing, for each item (i.e., attitude target), item-standardized implicit attitude scores on item-standardized predictions, and then averaging the resulting coefficients together. The results are presented in Table S1.

Study	Within-subjects correlation between predictions and implicit attitude scores	Between-subjects correlation between predictions and implicit attitude scores
1A	0.160	0.225
1B	0.157	0.226
2A	0.347	0.375
2B	0.303	0.335
2C	0.320	0.344
3A	0.410	0.496
3B	0.379	0.494

Table S1: Comparison of within- and between-subject correlations between predictions and implicit attitude scores for each study. (These correlations are not controlling for explicit attitudes.)

Overall, a comparison of the within-subjects effect sizes in the left column of Table S1 and the between-subjects effect sizes in the right column of Table S1 suggests that, unlike in past work by Hahn et al. (2014), the two types of analyses do not strongly dissociate. In fact, if anything, between-subjects analyses have produced bigger effects than did within-subjects analyses, suggesting that having to calibrate their responses did not interfere with participants' predictive accuracy in the present studies.

Supplementary Results from Studies 2B–2C and 3B

Plots relevant to the results reported below are available from the Open Science Framework (https://osf.io/vc7xp/).

Study 2B

Sources of Between-Attitude Variation in Predictive Accuracy. Predictive accuracy varied considerably as a function of the attitude object. When we regressed unstandardized D scores on unstandardized predictions for each attitude object separately, we found predictive accuracy ranging from $\beta = 0.035$ for the future/past IAT to $\beta = 0.653$ for the Democrats/Republicans IAT.

We also refit these regression models controlling for participants' explicit attitudes, providing us, for each attitude object, a measure of participants' accuracy at predicting the portion of their implicit attitudes that diverges from their explicit attitudes (or the "uniquely implicit" component of predictive accuracy). Participants' ability to predict the uniquely implicit component of their implicit attitudes varied widely across targets as well, with the poorest performance observed for the celebrities/regular people IAT ($\beta = -0.155$) and the best performance observed for the gun rights/gun control IAT ($\beta = 0.539$) when controlling for time-1 explicit attitudes, and the poorest performance observed for the nerds/jocks IAT ($\beta = -0.086$) and the best performance observed for the feminism/traditional values IAT ($\beta = 0.531$) when controlling for time-2 explicit attitudes.

Incremental relationships accounting for time-1 and time-2 explicit attitudes were significantly correlated with each other, $r = .756$, $t(58) = 8.80$, $p < .001$.

Explicit–Implicit Correlation. Next, we consider an additional variable that differs between attitude targets: the explicit–implicit correlation, or the extent to which implicit attitudes toward that target overlap with (or diverge from) explicit attitudes. We computed the explicit–implicit correlation for each of the 60 attitude targets, and then correlated this variable with participants' (unstandardized) predictive accuracy across attitude targets, separately with time-1 and time-2 explicit attitudes as the control variable.

Using time-1 explicit attitudes, the two variables were strongly positively related, $\beta = 0.748$, $t(58) = 8.58$, $p < .001$, indicating that participants tend to be successful at predicting the implicit attitudes that overlap most with their explicit attitudes. This relationship was strongly attenuated and nonsignificant when analyzing the uniquely implicit component of predictive accuracy as the dependent variable, $\beta = 0.218$, $t(58) = 1.70$, $p = .095$.

Using time-2 explicit attitudes, the two variables were also strongly positively related, $\beta = 0.852$, $t(58) = 12.39$, $p < .001$, indicating that participants tend to be successful at predicting the implicit attitudes that overlap most with their explicit attitudes. This relationship was strongly attenuated and nonsignificant when analyzing the uniquely implicit component of predictive accuracy as the dependent variable, $\beta = 0.078$, $t(58) = 0.60$, $p = .552$. Critically, time-1 and time-2 explicit attitudes produced inferentially equivalent results.

Discussion. Taken together, these results suggest that the findings reported in Study 2A are robust to the time of administering the explicit attitude measure (pre-prediction vs. post-prediction).

Study 2C

Sources of Between-Attitude Variation in Predictive Accuracy. Predictive accuracy varied considerably as a function of the attitude object. When we regressed unstandardized D scores on unstandardized predictions for each attitude object separately, we found predictive

accuracy ranging from $\beta = 0.048$ for the future/past IAT to $\beta = 0.629$ for the creationism/evolution IAT.

Participants' ability to predict the uniquely implicit component of their implicit attitudes varied widely across targets as well, with the poorest performance observed for the future/past IAT ($\beta = 0.031$) and the best performance observed for the creationism/evolution IAT ($\beta = 0.597$) when controlling for time-1 explicit attitudes, and the poorest performance observed for the gun rights/gun control IAT ($\beta = -0.082$) and the best performance observed for the religion/atheism IAT ($\beta = 0.525$) when controlling for time-2 explicit attitudes.

Incremental relationships accounting for time-1 and time-2 explicit attitudes were significantly correlated with each other, $r = .834$, $t(58) = 11.50$, $p < .001$.

Explicit–Implicit Correlation. Using time-1 explicit attitudes as the control variable, explicit–implicit correlations and predictive accuracy were strongly positively related, $\beta = 0.668$, $t(58) = 6.83$, $p < .001$, indicating that participants tend to be successful at predicting the implicit attitudes that overlap most with their explicit attitudes. This relationship was strongly attenuated and nonsignificant when analyzing the uniquely implicit component of predictive accuracy as the dependent variable, $\beta = 0.133$, $t(58) = 1.02$, $p = .311$.

Using time-2 explicit attitudes, the two variables were also strongly positively related, $\beta = 0.768$, $t(58) = 9.12$, $p < .001$, indicating that participants tend to be successful at predicting the implicit attitudes that overlap most with their explicit attitudes. This relationship was strongly attenuated and nonsignificant when analyzing the uniquely implicit component of predictive accuracy as the dependent variable, $\beta = 0.016$, $t(58) = 0.12$, $p = .903$. Critically, time-1 and time-2 explicit attitudes produced inferentially equivalent results.

Discussion. Taken together, these results suggest that the findings reported in Studies 2A–2B are robust to the time of administering the explicit attitude measure (pre-prediction vs. post-prediction) and the type of explicit measure used.

Study 3A

Demographic Predictability. As in Study 2A, we investigated whether participants showed more awareness for attitude targets that were more easily predictable from demographic information. The correlation between targets' demographic predictability and predictive accuracy was considerably smaller than in Study 2A and not statistically significant, $\beta = 0.19$, $t(14) = 0.72$, $p = .483$; however, this null result is difficult to interpret due to the considerably smaller sample of attitude objects. As before, the correlation dropped substantially when analyzing the uniquely implicit component of predictive accuracy, $\beta = 0.023$, $t(14) = 0.09$, $p = .932$, and participants still showed substantial accuracy for the subset of targets that were not predictable via demographic information, $\beta = 0.41$, $t(376.72) = 13.93$, $p < .001$.

Explicit–Implicit Correlation. Additionally, as in Study 2A, we tested whether participants were more accurate for attitude targets with higher explicit–implicit correlations. We computed the correlation between the explicit–implicit correlation and the prediction–implicit correlation for each attitude target (see Figure 9). When we used the zero-order relationships for each pair of variables, similar to Study 2A, the correlation of correlations was strongly positive, $\beta = .794$, $t(14) = 4.90$, $p < .001$, suggesting that participants tend to be successful at predicting their implicit attitudes when they overlap with their explicit attitudes. However, similar to Study 2A, when analyzing the uniquely implicit component of the prediction–implicit correlation (controlling for explicit attitudes), the correlation of correlations became considerably weaker and nonsignificant, $\beta = 0.12$, $t(14) = 0.45$, $p = .663$.

Study 3B

Sources of Between-Attitude Variation in Predictive Accuracy. Predictive accuracy varied considerably as a function of the attitude object. When we regressed unstandardized AMP scores on unstandardized predictions for each attitude object separately, we found predictive accuracy ranging from $\beta = 0.232$ for the children/adults AMP to $\beta = 0.636$ for the Biden/Trump AMP.

Participants' ability to predict the uniquely implicit component of their implicit attitudes varied widely across targets as well, with the poorest performance observed for the children/adults AMP ($\beta = 0.215$) and the best performance observed for the vegetables/meat AMP ($\beta = 0.540$) when controlling for time-1 explicit attitudes, and the poorest performance observed for the Tom Cruise/Denzel Washington AMP ($\beta = 0.130$) and the best performance observed for the Yankees/Diamondbacks AMP ($\beta = 0.564$) when controlling for time-2 explicit attitudes.

Incremental relationships accounting for time-1 and time-2 explicit attitudes were significantly correlated with each other, $r = .704$, $t(17) = 4.08$, $p < .001$.

Demographic Predictability. As in Study 3A, the correlation between attitude targets' demographic predictability and predictive accuracy was not statistically significant, $\beta = 0.397$, $t(17) = 1.78$, $p = 0.093$; however, this result is difficult to interpret due to the small sample of attitude objects. As before, the correlation dropped substantially when analyzing the uniquely implicit component of predictive accuracy, $\beta = -0.033$, $t(17) = -0.14$, $p = .892$, and participants still showed substantial accuracy for the subset of targets that were not predictable via demographic information, $\beta = 0.350$, $t(17.1) = 12.56$, $p < .001$.

Explicit–Implicit Correlation. Using time-1 explicit attitudes as the control variable, explicit–implicit correlations and predictive accuracy were strongly positively related, $\beta = 0.725$, $t(17) = 4.34$, $p < .001$, indicating that participants tend to be successful at predicting the implicit attitudes that overlap most with their explicit attitudes. This relationship was strongly attenuated and nonsignificant when analyzing the uniquely implicit component of predictive accuracy as the dependent variable, $\beta = 0.068$, $t(17) = 0.28$, $p = .782$.

Using time-2 explicit attitudes, the two variables were also strongly positively related, $\beta = 0.767$, $t(17) = 4.93$, $p < .001$, indicating that participants tend to be successful at predicting the implicit attitudes that overlap most with their explicit attitudes. This relationship was strongly attenuated and nonsignificant when analyzing the uniquely implicit

component of predictive accuracy as the dependent variable, β = -0.309, $t(17)$ = -1.34, p = .198. Critically, time-1 and time-2 explicit attitudes produced inferentially equivalent results.

Discussion. Taken together, these results suggest that the findings reported in Studies 2A–2C are robust to the type of implicit measure (IAT vs. AMP).

Sources of information underlying implicit attitude predictions: Details

In the main text, we regress participants' predictions on three potential sources of information – true implicit attitude scores, explicit attitudes, and demographic-based predictions – aggregating across Studies 1–3. Here, we repeat this analysis for each study separately. The results are shown in Table S2.

Study	Coefficient and statistics for implicit attitude scores	Coefficient and statistics for explicit attitude scores	Coefficient and statistics for demographic-based predictions
1A	β = 0.071, t(424.68) = 3.28, p = .001	β = 0.417, t(517.82) = 16.18, p < .001	β = 0.110, t(8.11) = 0.97, p = .358
1B	β = 0.059, t(394.43) = 2.74, p = .006	β = 0.440, t(501.75) = 16.77, p < .001	β = 0.345, t(4.13) = 2.91, p = .042
2A	β = 0.153, t(2268) = 16.16, p < .001	β = 0.586, t(75.70) = 50.97, p < .001	β = 0.039, t(54.14) = 1.83, p = .073
2B	β = 0.205, t(96.31) = 13.93, p < .001	β = 0.478, t(91.99) = 24.91, p < .001	β = 0.062, t(57.64) = 2.87, p = .006
2C	β = 0.220, t(791.26) = 13.09, p < .001	β = 0.421, t(77.83) = 18.78, p < .001	β = 0.064, t(57.81) = 2.43, p = .018

3A	$\beta = 0.174$, $t(26.92) = 5.67$, $p < .001$	$\beta = 0.558$, $t(44.11) = 20.08$, $p < .001$	$\beta = 0.389$, $t(1417) = 0.11$, $p = .914$
3B	$\beta = 0.270$, $t(334.87) = 11.74$, $p < .001$	$\beta = 0.446$, $t(26.28) = 15.76$, $p < .001$	$\beta = 0.117$, $t(13.82) = 3.02$, $p = .009$

Table S2: Information-source analysis for each study individually (reported in aggregate form in main text).

To summarize, the effect of all explicit and implicit attitudes was statistically significant in each study. The effect of demographic predictability was more variable and was significant in only four out of seven studies. Explicit attitudes consistently produced a larger effect than did demographic predictability or, critically, implicit attitudes, mirroring the combined results reported in the main text.

References

Hahn, A., Judd, C. M., Hirsh, H. K. & Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, *143*(3), 1369–1392. https://doi.org/10.1037/a0035028

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General*, *134*(4), 565–584. https://doi.org/10.1037/0096-3445.134.4.565