Pilot Testing

Word stimuli for IAT. Development and pre-testing of word stimuli were guided by procedures used in Amodio and Devine (2006). To develop the word stimuli for the intelligence IAT, the authors brainstormed a list of 28 "dumb" words and 13 "smart" words to examine implicit attitudes of intelligence. Thirty-five adults recruited from online forums (e.g., Craigslist, Reddit) and the university where the study took place rated the fit of each word with the respective categories on a scale ranging from 1 (Not at all) to 5 (Strongly). All words with a mean rating below the scale midpoint (i.e., 3) were removed. Additionally, words that included suffixes (e.g., brainless) or that could have more than one meaning (e.g., dense) were removed due to the additional cognitive processing needed to categorize the words. Seven of the remaining words from each category were piloted again.

In the second pilot, 57 adults recruited from online forums and the host university were asked if they were familiar with the definition of each word (*Yes/No*), the degree to which each word fit within the respective categories, and the degree to which each word is stereotypical of college students (non-athletes), Black student-athletes, and White student-athletes. Participants responded on a scale ranging from 1 (Not at all) to 5 (Strongly). One word ("math") was removed due to a mean fit rating below the scale midpoint ($M = 2.60$, $SD = 1.45$). Various groupings of five, six, and seven words were tested using paired samples *t*-tests until a grouping was found where no statistical differences in average rated fit was evident between "smart" and "dumb" word groups.[1] The final "smart" stimuli words were "brainy," "scholar," "educated," "scientist," and "academic." The final "dumb" stimuli words were "idiot," "stupid," "moron,"

[1] Paired samples t-test results indicated no statistical difference in fit between the final set of "Smart" ($M = 4.17$, $SD = 0.85$) and "Dumb" ($M = 4.00$, $SD = 1.01$) words, $t(56) = -1.256$, $p = .215$, $r = .37$.

"airhead," and "ignorant." *Smart* target words were rated as more stereotypical of non-athlete college students ($M = 2.94$, $SD = 0.59$) than of White student-athletes ($M = 2.39$, $SD = 0.84$), $t(54) = 5.575$, $p < .001$, and *Dumb* target words were rated as more stereotypical of White student-athletes ($M = 1.87$, $SD = 0.86$) than non-athlete college students ($M = 1.54$, $SD = 0.62$), $t(54) = -3.108$, $p < .01$. Finally, *Dumb* target words were rated as more stereotypical of White student-athletes ($M = 1.87$, $SD = 0.86$) than Black student-athletes ($M = 1.66$, $SD = 0.86$), $t(54) = -2.008$, $p = .049$, $r = .57$. No significant difference was found between stereotypical ratings of *Smart* target words between Black and White student-athletes.

To examine implicit attitudes of pleasantness in the BIAT, "pleasant" stimuli words, including "happy," "warm," "love," and "friend," were taken from Sriram and Greenwald's BIAT study (2009). To ensure these words were exemplar for use in the current study, they were tested in the second pilot study mentioned above. Participants were asked if they were familiar with the definition of the words in addition to the degree to which each of the words fit within the category "positive" on a scale ranging from 1 (Not at all) to 5 (Strongly). Almost all participants readily knew the definition of each word (only 1 participant indicated not readily knowing the definition of "love" and "friend"). On average, "positive" words were rated on the high end of the scale with regard to fit (mean range = 3.58–4.40), thus verifying that the words were exemplars for this category.

Photo stimuli for IAT. To develop the photo stimuli for the IAT, we conducted an extensive search of open access or public domain photos of male college students and student-athletes online. Images with neutral expressions were selected to represent eight Black student-athletes (BSA), eight White student-athletes (WSA), and eight White non-student-athletes (WNSA). No student-athletes from Division I schools were used due to the potential recognition

of these athletes by participants. Photos were modified using Adobe Photoshop. Individuals were placed on a solid maroon background reading "University of" across the top. The university name was intentionally cut off so as not to introduce potential participant bias associated with a particular university. All individuals were displayed from the waist up. "Student-athletes" appeared in a black football jersey with shoulder pads and the number "42" in grey. "College students (non-athletes) appeared in a black t-shirt designed to resemble common university apparel. Specifically, "University of" and "EST. 1885" appeared in gold lettering, and the university name was again cut off.

In the first round of pilot testing, participants recruited from online forums and the hosting university were asked to indicate the perceived age, gender, race/ethnicity, degrees awarded, grades in school, level of intelligence (compared to someone with "average" intelligence), pleasantness, and attractiveness of the individual in each photo. Each question utilized scaled response options with the exception of age which was open-ended. Pilot participants were also asked, "Is this person a college student?" and "Is this person an athlete?" (*Yes* or *No*). If the individual in the photo was identified as an athlete, participants were asked "At what level is this person an athlete?" [*High school, NCAA (college), Professional,* or *Other*]. Lastly, participants were asked to rate the overall quality of the photo on a 10-point scale ranging from "Very poor" to "Excellent." After each question, participants were also asked to report their confidence in their response using a slider scale ranging from 0% to 100%. Each pilot participant was randomly assigned to rate six of the 24 photos with the option to rate additional photos. Each photo was rated 13–36 times due to random assignment of photo blocks and some participants choosing to rate more than one block of photos. The first two authors reviewed preliminary analyses and excluded photos with greater numbers of identifiable problems (e.g., varying

responses regarding age/gender/race, very high or low intelligence/pleasant/attractiveness ratings, etc.). Two photos were removed from each category (WSA, BSA, WNSA) and the remaining eighteen photos were piloted again.

In the second round of pilot testing, pilot participants were asked to indicate the age, gender, race/ethnicity, pleasantness, and intelligence of the individual in the photo; college student-status of the depicted individual; and overall photo quality rating. Again, results were analyzed and examined for identifiable problems. One additional photo from each category was removed, leaving five photos per category. One-way repeated measures *ANOVA*s were conducted to ensure that there were not significant group differences in perceived intelligence or pleasantness between WSA and NSA (IAT 1/BIAT) and WSA and BSA (IAT 2) photo groups.[2]

[2] A one-way repeated measures *ANOVA* found statistical differences in intelligence between photo groups, $F(2,8) = 11.461$, $p = .004$. However, *post hoc* tests using the Bonferroni correction revealed no statistically significant differences in intelligence between WSA ($M = 4.13$, $SD = 0.08$) and NSA ($M = 4.35$, $SD = 0.16$) photos, $p = .107$. A one-way repeated measures *ANOVA* found no statistical differences in pleasantness between NSA ($M = 3.15$, $SD = 0.22$), WSA ($M = 3.23$, $SD = 0.15$), or BSA ($M = 3.23$, $SD = 0.15$) photo groups, $F(2,8) = 0.347$, $p = .717$.