**Excerpt from Supplement Materials of:**

Levari, D. E., Gilbert, D. T., Wilson, T. D., Sievers, B., Amodio, D. M., & Wheatley, T. (2018). Prevalence-induced concept change in human judgment. *Science*, *360*(6396), 1465-1467.

Methods for Study 7 Pretest

**Overview**. In Study 7, we asked participants to play the role of a reviewer on an IRB. We showed participants a series of proposals for scientific studies and asked them to decide whether researchers should be prohibited from conducting the study or should be allowed to conduct the study. The proposals varied in their ethicality. Over the course of many trials, we decreased the prevalence of unethical proposals for some participants. We predicted that these participants would respond to the decrease in the prevalence of unethical proposals by rejecting some proposals that were ethically identical to those they had previously accepted.

Whereas colors and computer-generated faces vary on physical continua that can be measured on a ratio scale, ethicality can at best be measured on an ordinal scale. As such, the materials and procedures for Study 7 differed somewhat from the materials and procedures used in our previous studies.

**Materials**. We wrote 381 short proposals for scientific experiments involving human participants. The proposals contained between 5 and 37 words (M = 25.34 words). We used our own judgment to preliminarily classify each proposal as either ethical, ambiguous, or unethical. We then recruited 361 U. S. residents (198 male, 161 female, 2 gender unspecified) via Amazon Mechanical Turk and asked them to read and rate a subset of these proposals. We will refer to these participants as *the raters*. Raters were told that (a) the proposals described experiments that were designed to be conducted with adults who had volunteered to take part in exchange for money; (b) all the studies described in the proposals were research on human behavior; (c) when scientists lie to participants either before or during a study, they always tell those participants the truth when the study is over; and (d) participants are always free to withdraw from a study at any time.

Each rater was paid $1 to read and rate 76 proposals. We divided the 381 proposals into a set of 15 proposals that were seen by all raters (the constant set) and a set of 366 proposals that were seen by a subset of raters (the variable set). Specifically, the 366 proposals were divided into 6 sets of 61 proposals (the variable sets), and each rater saw one of these 6 variable sets as well as the constant set of 15 proposals. Twenty-one of the proposals in each of the variable sets had been preliminarily classified as ethical, 23 had been preliminarily classified as ambiguous, and 17 had been preliminarily classified as unethical. The 61 proposals in each of the variable sets were presented in random order, and after the 20th and 40th, and 61st proposals we included a “catch question” to ensure that raters were reading carefully (viz., “If you're actually reading this question, please select the number 3 as your response. Thank you for reading all the questions carefully”). Each rater first saw one of the 6 variable sets of 61 proposals, and then saw the 15 proposals in the constant set. After seeing each proposal, raters were asked the question “Should this experiment be allowed to be conducted?” which they answered using a 7-point Likert scale whose endpoints were anchored with the phrases “Definitely not” (1) and “Definitely” (7). Raters spent between 3.18 and 53.72 minutes (M= 16.09 min) making their ratings. After they did so, raters completed several other measures including a Turing test (e.g., “If you’re reading this, type the word *banana*”), and supplied demographic information.

We excluded the ratings of two male and three female raters who failed the Turing test, and then computed the mean rating of each proposal. Despite the fact that participants’ ratings were inherently subjective, for the sake of consistency we will refer to the mean of each proposal’s ratings as its *objective ethicality.* Each rater saw 76 proposals. Fifteen of these proposals (the constant set) were seen by all raters, which allowed us to estimate how much the complete pool of raters agreed with regard to judgments of ethicality. Inter-rater reliability was quite high (Cronbach’s alpha = .85), indicating that raters were in very close agreement about the objective ethicality of the proposals. We used each proposal’s objective ethicality to classify it as a member of one of three categories. To ensure that we had a sufficient number of proposals in each of these categories, we classified proposals whose objective ethicality was greater than 6 and less than or equal to 7 as *ethical*; proposals whose objective ethicality was greater than 4 and less than or equal to 6 as *ambiguous*; and proposals whose objective ethicality was less than or equal to 4 and greater than or equal to 1 as *unethical*. We then selected the proposals in each of the three categories whose objective ethicality ratings had the lowest standard deviations. Specifically, we selected 113 *ethical* *proposals* (e.g., “Participants will make a list of the cities they would most like to visit around the world, and write about what they would do in each one”), 80 *ambiguous* *proposals* (“Participants will be given a plant and told that it is a natural remedy for itching. In reality, it will cause itching. Their reaction will be recorded”), and 80 *unethical proposals* (e.g., “Participants will be asked to lick a frozen piece of human fecal matter. Afterwards, they will be given mouthwash. The amount of mouthwash used will be measured”). These 273 proposals were used as materials in Study 7.