

Preregistration Report (08/03/2017)

Title: The A/B Illusion: Experiment 1 Exact Replication

Authors: Michelle N. Meyer, Patrick R. Heck, Geoffrey S. Holtzman, Christopher F. Chabris

Research Questions:

The A/B illusion is a hypothesized phenomenon in which individuals perceive the decision to run a randomized, controlled experiment (e.g., comparing two interventions, policies, or practices) on human subjects as less appropriate than simply implementing one of those alternatives without testing its effects. The A/B Illusion was previously anecdotally observed and described (Meyer, 2015), but it had never been experimentally investigated until our pilot research (see below). The research questions we are asking include:

1. Can we demonstrate the A/B illusion in naive research participants?
2. Assuming we are able to detect an effect, do any demographic variables or other individual differences either amplify or attenuate the A/B illusion?
3. What kinds of reasons do participants give for endorsing the A/B illusion, and what kinds of reasons do participants give for approving of unilateral implementation of untested policies?

Previous (Pilot) Research:

We ran a pilot experiment on 7/18/2017 which demonstrated that participants viewed the decision to implement a new healthcare policy (either placing informative posters in operations rooms or adding checklists to doctors' ID badges) as more appropriate than the decision to run a randomized controlled trial to evaluate the effectiveness of these two policies. A medium-sized, online sample ($N = 413$ Mechanical Turk workers) were randomly assigned to one of four conditions, each describing a specific policy implementation ("A" or "B") or a proposed 'experiment' to be run by a research director that would allow for an evaluation of the efficacy of each policy ("A/B"). An additional experimental condition was presented with a single additional sentence mentioning survival rates and plans to implement a universal policy in one year on the basis of the experiment's results. As predicted, the decision to implement a new policy (either A or B) was viewed as substantially more appropriate than the decision to run a controlled experiment testing the two policies against each other, regardless of whether the experimental condition mentioned survival rates or not. The observed effects were large (Cohen's d between 0.94 and 1.25), and they did not appear to differ substantially by preliminary demographic variables (sex, age, race, income). A PDF of this pilot survey will be uploaded alongside this preregistration document.

Comment boxes provided in the questionnaire of our pilot study revealed a number of cases in which participants misinterpreted or misunderstood the vignette. The most common misunderstanding was that the checklist was intended to inform patients (rather

than doctors) of the risks of the procedure or of safety precautions that the patients should take to avoid infection. However, the discovered effect was significant and large both when such participants were included in the analysis (as planned) and when they were excluded (for exploratory purposes). It is important to replicate this effect using identically-worded vignettes to ensure robustness and replicability of the A/B illusion, so that we may then conceptually replicate this effect in other domains in order to demonstrate generalizability.

Participants were also asked to provide their sex, race/ethnicity, age, educational attainment, and income, and to answer three questions about the relative importance of an interventional God, faith, feelings, and science. These additional variables did not differ between conditions and appeared not to affect the measure of interest (appropriateness).

The only differences between the pilot and the present study are the survey platform used to collect data and the day of the week and time of day of data collection. The pilot study used SurveyMonkey.com; this replication will use Qualtrics. Because of the way these platforms enable random assignment to conditions, we expect this replication to produce sample sizes much closer to equal across the four conditions, compared to the pilot study. All other wordings, incentives, and recruitment procedures will be carried over.

Hypotheses:

We predict that participants who read vignettes describing an unequivocal policy change decision (to either ‘Policy A’ - badges, or ‘Policy B’ - posters) will rate this decision as more appropriate than those participants who read a similarly unequivocal decision to run a randomized, controlled experiment designed to test the comparative effectiveness of ‘Policy A’ and ‘Policy B.’

Data Collection Procedures:

Sample & Sample Rationale.

Participants (≥ 18 years; restricted to the United States) will be recruited via Amazon Mechanical Turk. Participants will be paid \$0.25. We will recruit approximately 100 participants per vignette condition. The effect sizes observed in the pilot study ensure that samples of this size will adequately power our experiment ($> 95\%$ power).

Stimuli (Vignettes) and Additional Questions.

Badge

Some medical treatments require a doctor to insert a plastic tube into a large vein. These treatments can save lives, but they can also lead to deadly infections. A hospital director wants to reduce these infections, so he decides to give each doctor who performs this procedure a new ID badge with a list of standard safety precautions for the procedure

printed on the back. All patients having this procedure will then be treated by doctors with this list attached to their clothing.

How appropriate is the director's decision?

Poster

Some medical treatments require a doctor to insert a plastic tube into a large vein. These treatments can save lives, but they can also lead to deadly infections. A hospital director wants to reduce these infections, so he decides to hang a poster with a list of standard safety precautions for this procedure in all procedure rooms. All patients having this procedure will then be treated in rooms with this list posted on the wall.

How appropriate is the director's decision?

BPS

Some medical treatments require a doctor to insert a plastic tube into a large vein. These treatments can save lives, but they can also lead to deadly infections. A hospital director thinks of two different ways to reduce these infections, so he decides to run an experiment by randomly assigning patients to one of two test conditions. Half of patients will be treated by a doctor who has received a new ID badge with a list of standard safety precautions for the procedure printed on the back. The other half will be treated in a room with a poster listing the same precautions hanging on the wall.

How appropriate is the director's decision?

BPL

Some medical treatments require a doctor to insert a plastic tube into a large vein. These treatments can save lives, but they can also lead to deadly infections. A hospital director thinks of two different ways to reduce these infections, so he decides to run an experiment by randomly assigning patients to one of two test conditions. Half of patients will be treated by doctors who have received a new ID badge with a list of standard safety precautions for the procedure printed on the back. The other half will be treated in rooms with a poster listing the same precautions hanging on the wall. After a year, the director will have all patients treated in whichever way turns out to have the highest survival rate.

How appropriate is the director's decision?

Primary Dependent Measure: Participants will be asked to rate the decision on a 1-5 scale measuring appropriateness. Participants will also be asked to complete a free-response item asking them why they chose the rating they gave.

Additional items: We will ask participants to provide their sex, age, income, race/ethnicity, educational attainment, three questions about the relative importance of God, faith, feelings, and science, and science literacy/education.

Data Exclusions.

We intend to exclude the following participants:

1. Those who demonstrate a clear misunderstanding of the vignette, as determined by their free-response explanation. An example of a response that excluded a participant under this criterion in the pilot study is a participant in an AB condition who stated “I believe it would give a detailed POV on how some patients would react to the readily available info.” (The interventions are designed to remind doctors, not inform patients.)
2. Those whose free-response explanation makes clear that they did not take at least that task seriously. Examples of responses that excluded participants under this criterion in the pilot study include: “meow meow meow,” “blah blah blah,” and “fdff.”
3. Those whose appropriateness ratings are at odds with their free-response explanation. Examples of responses that excluded participants under this criterion in the pilot study include a participant assigned to the Policy A – Badge condition who stated “I think putting extra emphasis on prevention is a good thing, it would give doctors something to reference so that steps don’t get skipped when they are busy,” but who rated the director’s decision as “very inappropriate”; another participant assigned to the Badge condition who stated “Because somethign [sic] slightly tedious will save lives, it is worth it.,” but rated the director’s decision “somewhat inappropriate”; and a participant assigned to the Poster condition who stated “It could be a matter of life and death, it should be taken seriously. The poster with [sic] help medical workers be reminded of that.,” but who rated the director’s decision “very inappropriate”. We suspect that these participants confused the poles of the scale.
4. Participants who took part in the pilot study or in any of our other studies in this project, as determined by their MTurk IDs.

No other exclusions are planned.

Planned Analyses:

Critical (Step 1) analyses:

Independent groups *t*-test for mean-level differences in appropriateness between A [badge] and B [poster]. This is just for descriptive purposes.

Independent groups *t*-test for differences in appropriateness between the two AB conditions [bp_long] and [bp_short]. This is just for descriptive purposes.

Compare A [badge] appropriateness with the two AB conditions [bp_long] and [bp_short] using two separate independent groups *t*-tests. The AB Illusion hypothesis predicts that A [badge] will be rated more appropriate than either AB case.

Compare B [poster] appropriateness with the two AB conditions [bp_long] and [bp_short] using two separate independent groups *t*-tests. The AB Illusion hypothesis predicts that B [poster] will be rated more appropriate than either AB case.

To obtain an estimate of overall effect size *d*, we will compare the magnitude of the difference in appropriateness between [A] and [B] combined and [AB_long], and the difference between [A] and [B] combined and [AB_short].

Belief in intuition, God, and science (Step 2) analyses:

In each of the four conditions, regress appropriateness on the intuitive beliefs variable (answers to the three questions querying beliefs in God, intuition, and science). The pilot study determined that these questions were not predictive of appropriateness ratings; we decided to use them again to make this experiment an exact replication.

Exploratory Demographic (Step 3) analyses:

We will conduct exploratory demographic analyses on: sex, race/ethnicity, age, education, and income, to determine whether any subgroup of participants judges the experimental (A/B) conditions as more or less appropriate than other subgroups do.

Free Response Question (Step 4) analysis:

A trained coder blind to participant condition and responses will employ a coding scheme to mark the free-response descriptions as indicated in the codebook below. Individual responses will be coded with multiple labels if they fit the criteria for more than one category.

We hypothesize that, among respondents in all conditions who rate the decision as somewhat or very appropriate, the most common coding of their free responses will be *Benefit*.

We hypothesize that, among respondents assigned to the A and B conditions who rate the decision as somewhat or very inappropriate, the most common coding of their responses will be *Ineffective*.

We hypothesize that, among respondents in the AB conditions who rate the decision as somewhat or very inappropriate, the most common coding of their free responses will be *Negative Research*.

We hypothesize that more respondents in the AB conditions will object to the apparent lack of patient consent (meeting the conditions for coding their free responses as *Consent*) than will respondents in the A and B conditions.

Reasons Given for (In)Appropriate Ratings in all Conditions: Codebook

1. Benefit

- | | |
|---------------------|--|
| <i>1.1 Benefit</i> | Indication that director's intervention (badges, posters, or experiment comparing badges and posters) will or might be effective in reducing infections or helping patients. |
| <i>1.2 Learning</i> | Specific mention that the intervention will or might help the director learn what will work or what will work best, that it will produce needed evidence, etc. (may or may not include specific positive mention of randomization or experimentation). |

2. No harm

- | | |
|------------------------------|--|
| <i>2.1 Absence of harm</i> | Comment that the respondent believes the intervention won't or is unlikely to do any harm. |
| <i>2.2 Positive equality</i> | In the A or B conditions, a comment that all patients are being treated the same way. In an AB condition, a comment that the two different groups are actually more or less the same (e.g., patients receive the same treatment, which is what counts; or all doctors receive the same informational reminders, just displayed differently). |

3. Harm

- 3.1 Medical risk/harm* Comment that the intervention will or may place some or all patients at medical risk or will or may medically harm some or all of them (e.g., because half of patients may or will receive an inferior intervention leading to greater infection rates or because doctors handling badges to review safety procedures may or will compromise a sterile environment).
- 3.2 Other risk/harm* Comment that the intervention will or may place some or all patients at non-medical risk or will or may harm some or all patients in some non-medical way (e.g., by causing patients anxiety about whether their doctor is competent).

4. No benefit

- 4.1 Ineffective* Comments that the intervention won't be, or is unlikely to be, effective in achieving the goal of reducing infections or helping patients (e.g., "won't work," "no point," "redundant" due to prior training).

5. Negative research

- 5.1 Randomization* Negative comments about randomization as a methodological approach. May include sound or unsound research method, sample size, reduced bias, dangers of randomization, concerns about study design (must clearly address randomization, either by name or proxy [i.e., "gold standard"]).
- 5.2 Experimentation* Negative comments about experimentation, testing, research, or studies, including (for negative comments) "guinea pigs," "lab rats," "playing with lives," "gambling with lives," "playing God," or wanting control over health care or medications.

5.3 Negative inequality

Negative comment that patients will be treated differently or unequally.

6. Consent

6.1 Notice

Comment on the importance of telling patients about the intervention, criticism of the apparent failure to disclose the intervention to patients, etc.

6.2 Consent

Comment on the importance of patient choice to participate or not in the intervention, criticism of the apparent failure to obtain patient consent to the intervention, etc.

7. Action

7.1 Act now

Comment that the director should act immediately rather than conducting an experiment, or that conducting the experiment for one year before making a decision is too long.

7.2 Best judgment

Comment that the director should “just use his best judgment” or that he should “just do what works best” instead of running an experiment.

8. Intent

8.1 Good intentions

Comment that the director’s intentions are good.

8.2 Bad intentions

Comment that the director’s intentions are bad.

9. Status quo

- 9.1 Status quo* Comment that the appropriateness of the director’s decision depends on how things (e.g., safety reminders to doctors) are currently done, or that all patients should receive “standard of care,” or words to that effect.

10. Other

- 10.1 Misunderstandings* Comments that reveal misunderstandings of the vignette (e.g., checklists designed to inform patients rather than remind doctors).
- 10.2 Irrelevant, unclear, other* Comments that are irrelevant (e.g., “meow meow mewo”) or insufficiently clear to interpret, or that make substantive points but do not fit any of the above categories.

Reference

Meyer, M. N. (2015). Two Cheers for Corporate Experimentation: The A/B Illusion and the Virtues of Data-Driven Innovation. *Colorado Technology Law Journal*, 13(2), 273-331.