Critical Condition: People Don't Dislike A Corporate Experiment More than They Dislike Its Worst Condition

Robert Mislavsky
Carey Business School
Johns Hopkins University
mislavsky@jhu.edu

Berkeley Dietvorst
Booth School of Business
University of Chicago
berkeley.dietvorst@chicagobooth.edu

Uri Simonsohn
ESADE
Ramon Llull University
uri.simonsohn@esade.edu

Abstract:

Why have companies faced a backlash for running experiments? Academics and pundits have argued people find corporate experimentation intrinsically objectionable. Here we investigate "experiment aversion," finding evidence that, if anything, experiments are *more* acceptable than the worst policies they contain. In six studies participants evaluated the acceptability of either corporate policy changes or of experiments testing them. When all policy changes were deemed acceptable, so was the experiment, even when it involved deception, unequal outcomes, and lack of consent. When a policy change was deemed unacceptable, so was the experiment, but less so. The acceptability of an experiment hinges on its critical condition—its least acceptable policy. Experiments are not unpopular, unpopular policies are unpopular.

Keywords: field experiments, public opinion, market research, business ethics

Acknowledgements:

We thank the review team for their constructive comments. We also thank Laura Kuder, Johanna Matt-Navarro, and Catherine O'Donnell for valuable research assistance and the Wharton Risk Management and Decision Processes Center, the Wharton Behavioral Lab, and the University of Chicago Booth School of Business for financial support.

In June 2014, the Proceedings of the National Academy of Science (PNAS) published an article describing the results of a field experiment where academic authors (Kramer, Guillory, & Hancock, 2014) partnered with Facebook to manipulate content users saw (i.e., "News Feeds"), showing either more positive or more negative emotional content, to measure potential emotional contagion. A month later, the online dating site OkCupid published a blog post titled "We Experiment on Human Beings," which described three experiments they had run on their users (Rudder, 2014). Reaction to the revelation of these experiments was swift and highly negative.

The backlash the Facebook and OkCupid experiments received, described by a Forbes contributor as "one epic freak out" (Muse, 2014), dominated several news cycles despite competing for attention with the 2014 World Cup and major U.S. Supreme Court rulings. Articles describing the negative reaction to the Facebook experiment reached the front page of the Wall Street Journal and were the number one most popular/shared articles on several news outlets, including The Atlantic, The Wall Street Journal, and The BBC.[1] Articles on CNN.com and in the New York Times proclaimed that Facebook treated users like "lab rats" (Goel, 2014; Goldman, 2014). When the OkCupid experiment was revealed, an article in FastCompany declared that the experiment was "way creepier" than Facebook's (Greenfield, 2014). Even legislators got involved, calling for investigations into data collection practices (R. Meyer, 2014; Stampler, 2014). A few months later, Facebook's chief technology officer formally acknowledged that the company was "unprepared" for the reaction elicited by the experiments and admitted that they "should have considered non-experimental ways" to conduct research on the topic (Schroepfer, 2014).

[1] Screenshots from the cited media coverage available from http://osf.io/z39aq.

In this paper, we present evidence suggesting that the backlash to these experiments was likely driven by the specific policies that these experiments contained (i.e., the individual treatment arms), rather than the fact that the policies were implemented as part of an experiment. As a result, we posit that reactions would have been at least as negative if these treatments were implemented as standalone policy changes, outside of an experimental context. We conclude that marketing researchers and organizational decision makers face similar scrutiny for running experiments as they do for implementing policy changes. Thus, we propose that organizations will not face more backlash when they implement objectionable policies as part of an experiment. Similarly, we propose that implementing objectionable policies outside of an experiment will not make them more palatable to the public.

FIELD EXPERIMENTS AND MARKETING SCIENCE

Experimentation provides an unrivalled source of actionable intelligence for businesses, governments, and non-profit organizations (Zoumpoulis, Simester, & Evgeniou, 2015), allowing researchers to identify the causal effects that alternative policies have on behavior.[2] Field experiments overcome the lower external validity of stylized lab experiments by taking place in the precise environment where specific policy changes will occur (DellaVigna, 2009). In part because of these advantages, field experimentation has become a popular tool for marketing scholars that is used to test and complement existing theory, as well as develop new insights into buyer behavior on wide-ranging topics. In this journal alone, field experiments have been used to explore charitable giving behavior (Sudhir, Roy, & Cherian, 2016), the effect of social influence on the adoption of new technologies (Miller & Mobarak, 2014), strategies for inducing multi-

[2] We define an experiment as an instance where an organization implements, at random, different policies for different groups with the intention of learning how they differently influence a specific outcome.

channel buying (Montaguti, Neslin, & Valentini, 2015), and consumer purchasing habits after the end of a promotion (Wang, Lewis, Cryder, & Sprigg, 2016).

Given the value of field experimentation, concerns about its acceptability must be taken seriously. Many pundits and scholars have interpreted the backlash to well-known field experiments as evidence that people have a broad and substantial aversion to experimentation. Gino (2015), for instance, proposed that managers are hesitant to run experiments within their own organizations, in part because they believe that customers and employees do not want to be experimented on. Hill (2014) found that companies that do run experiments often resort to using terms like "diagnostic test" or "A/B test" to avoid presumed negative associations with experimentation (see also, Luca, 2014). Meyer (2015) stated that people view field experiments as "more morally suspicious than an immediate, universal implementation of an untested practice" (p. 278) and titled this preference the "A/B illusion."

If consumers are indeed averse to experimentation, it would constitute an important barrier to evidence-based marketing and future collaborations between academics and organizations. Organizational decision makers may hesitate to run or publicize the results of experiments for fear of negative publicity, and customers may fear engaging with companies that they believe will experiment on them. In this article, we investigate whether or not such an aversion to experimentation exists.

THREE FORMS OF EXPERIMENT EVALUATION

We define three different forms that experiment evaluation could take and then preview our ability to empirically distinguish among them in this article:

1. *Absolute* experiment aversion – All experiments are deemed unacceptable, independent of the policies they include.

2. *Relative* experiment aversion – An experiment is less acceptable than the policies it contains, either because experimentation is a negative attribute (i.e., a negative main effect of experimentation), or because the underlying policies are deemed less acceptable when they are part of an experiment (i.e., an interaction between experimentation and policy acceptability). This means that experiments with acceptable policies could still be considered acceptable in absolute terms, but less acceptable than their underlying policies. Similarly, an experiment with an unacceptable policy could be viewed more negatively than the unacceptable policy on its own.
3. *Critical condition* – There is no experiment aversion. The acceptability of an experiment is instead a weighted average of the acceptability of its policies. Most importantly, this implies an experiment is no less acceptable than its least acceptable policy. Therefore, if an experiment is viewed negatively, it is only because one (or more) of its conditions (i.e., its "critical" condition) is viewed negatively and not because experimentation is a negative attribute *per se*. People will find an experiment that contains only acceptable conditions to be acceptable.

In Studies 1 and 2, we test for absolute experiment aversion and find several instances where experiments are, in fact, rated positively. Thus, we reject absolute experiment aversion. In Studies 3 and 4, we directly pit the acceptability of experiments against the acceptability of their underlying policies. Consistent with the *critical condition* account of experiment evaluation, we find that experiments are rated as no less acceptable than their least acceptable policies. Experiments, however, were also rated as less acceptable than the simple average acceptability of their underlying policies. This may reflect either moderate relative experiment aversion or

negativity bias, where people give more weight to negative attributes than to positive ones (e.g., Folkes & Kamins, 1999; Rozin & Royzman, 2001; Skowronski & Carlston, 1989). In Study 5 we tease these two apart by asking participants to evaluate experiments with two positive policies that are similarly acceptable (and where negativity bias should be absent). We find no evidence of even modest experiment aversion. In Study 6, we also include two similarly negative arms, again finding no evidence of relative experiment aversion. In sum, our evidence is inconsistent with both absolute and relative experiment aversion, and consistent with the "critical condition" account of experiment aversion.

TRANSPARENT REPORTING

In all 6 studies, participants read scenarios describing an action that a company could take (either an experiment or a universal policy change) and indicated how acceptable each action is. We ran all studies, except for Studies 3b and 6, on Amazon's Mechanical Turk (MTurk) using Qualtrics. Study 3b was a pen-and-paper survey of non-academic university staff. Study 6 was run in collaboration with Lucid (http://luc.id), a market research firm.

Study materials, data, analysis code, and supplements for all studies as well as pre-registrations for Studies 3b-6 are available at http://osf.io/z39aq. We report studies in the order they were conducted (except for Study 3b, which was added at the request of reviewers and conducted after Study 4) and discuss all additional studies conducted but not reported in the paper in Supplements 6 and 7. For all studies, we determined sample size before beginning data collection.[3] We report all data exclusions, all manipulations, and all measures.

[3] In our online studies, we typically obtained sample sizes that slightly exceeded our goals because some participants did not submit a completion code, allowing additional participants to take the survey. Participants, identified by their MTurk ID number, were not able to participate in more than one study. We included an attention check (Oppenheimer, Meyvis, & Davidenko, 2009) in the first question, and only those who answered correctly were able to participate in the studies. All participant responses are included in analyses, regardless of whether or not they completed the entire survey.

STUDY 1: PEOPLE DO FIND (SOME) EXPERIMENTS ACCEPTABLE

Our first study tests for absolute experiment aversion—people always object to experiments, even if all conditions are unambiguously beneficial. We presented participants with descriptions of corporate experiments that contained unambiguously positive conditions (e.g., giving $5 to employees for visiting the gym) or unambiguously negative conditions (e.g., taking $5 from employees for not visiting the gym). If absolute experiment aversion exists, participants should find all experiments objectionable. If experiments are instead evaluated based on their conditions, participants should only object to experiments that contain unambiguously negative conditions. Throughout these scenarios, we also added various aspects of experimentation that may contribute to experiment aversion, such as deception and lack of consent. If these specific features cause experiment aversion, participants should view these experiments negatively, even if they have only unambiguously positive conditions.

Method

Sample. We recruited 577 participants on MTurk, of which 505 successfully passed the attention check (37.5% female, M_{age} = 34.1 years). Participants were paid $0.75 for completing the study.

Design. Participants were assigned to one of ten experimental conditions. Fifty-three participants were assigned to the *policy change* condition. The remaining participants (N = 452) were assigned to one of nine *experiment* conditions.

Participants in the *policy change* condition read descriptions of nine possible policy changes. These involved *bad*, *good* or *very good* outcomes, in three different contexts. See

Table 1. Participants evaluated all nine policies in random order, answering three questions about their acceptability. We average them (Cronbach's $\alpha = .96$) to construct the "policy acceptability index." These ratings served as a manipulation check for our stimuli in the *experiment* conditions.

Participants in the nine *experiment* conditions read one scenario about a company running an experiment that randomly assigned employees/customers to one of two policy changes from one of the three contexts in Table 1. The condition pairs were *bad/good*, *control/good*, or *good/very good*. For example, the *shipping control/good* scenario read:

> "A shopping company runs an experiment on their shipping system where one group of customers is randomly picked and the company starts upgrading all 'Standard 5-day' shipped packages to 'Priority 3-day' shipping (without changing the cost to the customer). Another group of customers is randomly picked and gets no change in their shipping. The company will then compare customer satisfaction across the two groups."

Participants then answered the same three questions from the *policy change* condition (measures 1-3 in Table 1), but now focusing on the experiment as a whole rather than the underlying policies. They also answered three additional questions designed to more unambiguously evaluate the acceptability of the experiment (rather than willingness to participate in it). We average only these additional three questions ($\alpha = .86$) to construct the "experiment acceptability index." [4]

INSERT TABLE 1 ABOUT HERE

---

[4] In hindsight we found questions 1-3 to be ambiguous for interpreting the evaluation of experiments. Therefore, the experiment acceptability index in the main text is based only on questions 4-6. We report results aggregating over all 6 questions in footnote 5.

Participants also answered five comprehension checks to ensure they noticed potentially controversial attributes of the experiments (e.g., "People will be included in this study without agreeing to be included"). No other measures were collected in this condition. Results for measures not reported below are reported in Supplement 1.

Results

Acceptability of policy changes. Validating our choice of stimuli, the overall policy acceptability index for *bad* policy changes (M = 1.91) was below the midpoint (4) and below both the *good* (M = 6.18) and *very good* policy changes (M = 6.11), which were both above the midpoint. All t-tests vs. midpoint are ts > 20.9, *p*s < .001. The *good* and *very good* policies were rated as similarly acceptable, t(312) =.48, $p = .63$, and were close to the highest possible rating (medians of 6.7 and 7 respectively, on a 7-point scale).

Acceptability of experiments. Figure 1 shows the average *experiment acceptability index* for the nine *experiment* conditions. The results are inconsistent with absolute experiment aversion. In particular, when experiments did not include an objectionable condition (*control/good*, M = 5.11; *good/very good,* M = 5.17), they were rated above the midpoint and as more acceptable than when experiments did include an objectionable condition (*bad/good*, M = 3.25). The experiments with objectionable conditions were in turn rated below the midpoint. All t-tests vs. midpoint are ts > 5.9, *p*s < .001. People found experiments to be acceptable when all conditions in the experiment were acceptable and found experiments to be unacceptable when a condition in the experiment was unacceptable.[5]

Discussion

[5] These results are based on questions 4-6 in Table 1 (see footnote 4). Including all six questions, the results are very similar. Experiments with a bad condition (*bad/good*, M = 3.01) were rated below the midpoint and below experiments without a bad condition (*control/good*, M = 5.17; *good/very good*, M = 5.30), which were both above the midpoint. All t-tests vs. midpoint are ts > 8.4, *p*s < .001.

The results from Study 1 are inconsistent with absolute experiment aversion and consistent with a critical condition account of experiment evaluation. Additionally, participants found experiments with deception (e.g., one shipping speed was promised, another was actually delivered), unequal outcomes (e.g., some participants get $5 for attending the gym, others get $10), and lack of consent, to be acceptable, as long as all conditions were themselves acceptable.

INSERT FIGURE 1 ABOUT HERE

However, Study 1 has some important limitations. First, the experiments evaluated as acceptable had unambiguously beneficial outcomes (e.g., free shipping upgrade) and may not generalize to more routine corporate experiments where benefits to participants, if any, are less obvious. Second, we measured agreement with statements rather than absolute measures of acceptability, making it difficult to know whether the experiments are sufficiently acceptable. For example, the *good/very good* experiments were rated M = 5.17 on a 7-point scale where 7 implies strong agreement with the experiment being acceptable. While this is significantly above the midpoint, is it high enough to suggest people would not object to the experiment? Third, participants' ratings in the *policy change* and *experiment* conditions are not directly comparable because: (i) the sets of dependent variables, and their interpretation, are different in the *policy* and *experiment* conditions (see footnotes 4 and 5) and (ii) participants saw all nine policies in the *policy change* condition and only two in the *experiment* conditions. Fourth, participants in this

experiment may have higher than average tolerance for experiments because they routinely volunteer for experiments on Amazon Mechanical Turk.

In Studies 2-6 we address all of these issues. We use a wider variety of stimuli (Studies 3a and 3b) and have participants evaluate experiments similar to (controversial) experiments that companies have actually run (Studies 2 and 4). We use questions with less ambiguous endpoints (Studies 2-6) and with neutral and labeled midpoints (Studies 3-6). We have participants in the *policy change* condition rate only the two policy changes that are included in the corresponding *experiment* condition (Studies 3-6) and use the same measures of acceptability across conditions (Studies 2-6). Finally, in Studies 3b and 6, we recruited participants who do not routinely volunteer for experiments.

INSERT TABLE 2 ABOUT HERE

STUDY 2: PREDICTING EXPERIMENT RATINGS FROM CONDITION RATINGS

Kramer et al. (2014) ran an experiment studying emotional contagion through social networks. They manipulated mood by modifying the emotional content of Facebook users' status updates and measured its effect on users' subsequent emotion expression, which upset many users and spurred public outrage (Albergotti, 2014). If, as we have conjectured, people objected to the study because of its polices and not just because it was an experiment, then they should not object to a similar experiment with only acceptable conditions. In Study 2a, we conduct an exploratory search for acceptable and unacceptable mood inductions Facebook could have employed. In Study 2b, we test if the acceptability of the experiment hinges on the acceptability of the mood inductions used.

STUDY 2A: FINDING (UN)ACCEPTABLE MOOD INDUCTIONS

Method

Sample. We recruited 382 participants on MTurk, of which 303 passed the attention check (40.7% female, M_{age} = 30.3 years). Participants were paid $0.30 for completing the study.

Design. We generated six interventions, involving positive and negative versions of three possible changes to the site—showing only sad ads, showing only happy ads, showing sad status updates first, showing happy status updates first, showing the least liked status updates first, and showing the most liked status updates first. Each participant evaluated three alternative policies, one for each possible change to the site, randomizing whether participants saw the positive or negative change. We counterbalanced the order of the stimuli.

Measures. Participants answered two questions for each policy change: "Is it okay for a company to do this?" and "Would you object to a company doing this?" These questions were answered on 7-point scales, with endpoints labeled "1. It's definitely not okay"/ "7. It's definitely okay" and "1. I would definitely not object"/ "7. I definitely would object," respectively. We average the two items (r = 0.69; second question reverse-coded) to construct the "policy acceptability index."

Results

Participants found negative changes less acceptable than positive ones and manipulating status updates less acceptable than manipulating ads. From most to least acceptable, they ranked happy ads (M = 5.67), most liked status updates (M = 4.63), happy status updates (M = 4.58), sad ads (M = 3.90), least liked status updates (M = 3.62) and sad status updates (M = 3.08). For Study 2b, we used the highest rated (happy ads) and lowest rated (sad status updates) changes to

test our prediction that experiments are only objectionable if they contain objectionable conditions.

STUDY 2B: EXPERIMENTS WITH (UN)ACCEPTABLE MOOD INDUCTIONS

Method

Sample. We recruited 255 participants on MTurk, of which 201 passed the attention check (43.9% female, M_{age} = 34.2 years). Participants were paid $0.30 for completing the study.

Design. Participants were randomly assigned to one of two conditions in a between-subjects design. In both conditions, participants read descriptions of a social networking company that ran an experiment, assigning half of its customers to a control condition and the other half to a treatment condition. The treatment condition in those experiments was either the *happy ads* or *sad status updates* policy described in Study 2a. Participants answered the same two acceptability questions from Study 2a.

Results

The results were consistent with the critical condition account of experiment evaluation and inconsistent with absolute experiment aversion; only the experiment with an objectionable condition was considered objectionable. Participants rated the *happy ads* experiment significantly above the midpoint (M = 4.72), t(98) = 3.47, $p < .001$, and the *sad status updates* experiment below it (M = 2.59), t(99) = 9.30, $p < .001$.

Although Study 2 shows that experiments with acceptable conditions are acceptable in an absolute sense, relative experiment aversion may still exist if experiments are rated as being *less* acceptable than their underlying conditions. In Study 3 we examine this possibility by directly comparing ratings of individual policies to experiments that use these policies as conditions.

STUDY 3A: TESTING FOR RELATIVE EXPERIMENT AVERSION

Method

Sample. We recruited 533 participants on MTurk, of which 423 passed the attention check (43.5% female, M_{age} = 36.0 years). Participants were paid $0.50 for completing the study.

Design. Participants were randomly assigned to one of six conditions, in a 2 (action: *policy change* vs. *experiment*) x 3 (policy combination: *negative/positive* vs. *no change/positive* vs. *negative/no change*) fully between-subjects design.

Participants in the *policy change* conditions were told that a company was deciding between two policies. They were then told to imagine the company chose one of the policies and answered three questions about the acceptability of this action. They then answered the same questions, but imagining that the other policy had been chosen.

Participants in the *experiment* conditions were told that a company was running an experiment that randomly assigned customers to one of two policies (from the same pool of policy pairs as the *policy change* conditions) and answered the same questions as the *policy change* conditions.

Stimulus selection and sampling. To reduce the probability that the results would be driven by idiosyncratic features of the selected stimuli (Wells & Windschitl, 1999), we presented policy changes for seven different contexts (e.g., showing emotionally charged ads, changing a product recommendation system, and changing frequency of issuing coupons). See Supplement 2 for a full list of stimuli.

Measures. Participants in all conditions answered the following three questions containing labeled neutral midpoints:

1. How okay is it for the company to do this?
   (1 = It's really bad; 4 = It's okay; 7 = It's really good)

2. If you were a customer of this company and learned about the company's plans, how would this influence your opinion of the company?
   (1 = I would view the company much more negatively; 4 = […] not view the company any differently; 7 = […] much more positively)
3. If you were a customer of this company and learned about the company's plans, how likely would you be to switch to a different company?
   (1 = […] definitely not switch […]; 4 = […] not change how likely I am to switch [...]; 7 = […] would definitely switch […]; reverse-coded)

Participants in the *policy change* condition answered these questions twice, once for each policy (in counterbalanced order). Participants in the *experiment* condition answered these questions once, evaluating only the experiment. We average these items ($\alpha = .86$) to construct an "acceptability index."

Results

Evaluating policy changes. Validating our choice of stimuli, the *negative* policies were rated as the least acceptable (M = 2.65), followed by the *no change* (M = 4.61) and *positive* (M = 5.46) policies. The *negative* policies were rated below the midpoint (4), while the *no change* and *positive* policies were rated above the midpoint, all ts > 6.4, *p*s < .001.

Evaluating experiments. Replicating the results from Studies 1 and 2, and again inconsistent with absolute experiment aversion, experiments that only included acceptable policy changes (*no change/positive*) were rated as acceptable (M = 4.38); significantly above midpoint, t(79) = 3.54, $p < .001$. Conversely, experiments with an unacceptable policy (*negative/positive,* M = 3.22; *negative/no change,* M = 3.31) were rated below the midpoint, ts > 4.3, *p*s < .001. Because, in this study, we used a labeled neutral midpoint (see '*Measures'* above), evaluations above/below the midpoint are unambiguously positive/negative.

Because participants may not all have the same opinion of which policy is "worst," we compare participants' average ratings of each experiment in the *experiment* conditions to the

average rating of each participant's less preferred policy in the corresponding *policy change* conditions. When comparing average experiment ratings to the average of the lowest rated corresponding policies, participants found experiments to be significantly more acceptable in the *no change/positive*, $t(139) = 2.53$, $p = .013$, and *negative/positive*, $t(139) = 4.23$, $p < .001$, conditions, and marginally more acceptable in the *negative/no change* conditions, $t(137) = 1.77$, $p = .079$. Collapsing across all policy combinations, experiments were rated as significantly more acceptable than the policy that represented their least acceptable condition, $t(419)=5.16$, $p < .001$.[6] Most importantly, experiments were not rated as less acceptable than their worst conditions (see Figure 2). This suggests that participants rate experiments as some weighted average of its policies.

INSERT FIGURE 2 ABOUT HERE

STUDY 3B: REPLICATION WITH FIELD SURVEY

One concern about the generalizability of our findings may be that our results to this point have relied on a sample (MTurkers) that regularly opts-in to taking experiments and may therefore be less experiment averse than the general public. In this study, following suggestions of the review team, we replicated our findings using a sample of participants from outside an established participant pool.

[6] These results are consistent when comparing each experiment to the policy change with the lowest average rating (as opposed to the average of each participant's lowest rated policy). Experiments were rated directionally more acceptable than their worst policies in all three cases (significantly so for the negative/positive experiment; $t(139) = 3.94$, $p < .001$, negative/no change experiment, $t(132) = 2.06$, $p = .04$, and when collapsing across all policy pairs, $t(419) = 4.27$, $p < .001$).

Method

Sample. Three research assistants walked around a university campus, approached non-academic staff members, and asked them if they were willing to take a short, one-page pen-and-paper survey. We specifically instructed the research assistants to approach staff in and around non-academic buildings (e.g., the student union and library) to reduce the likelihood that our participants themselves would be involved in conducting research. It is also important to note that our respondents did not initiate participation in the study (reducing potential selection effects), nor were they compensated for completing the survey (which may have caused them to view academic research and experimentation more favorably). In total, we obtained 247 responses (68.4% female, M_{age} = 33.4 years).

Design. Participants were assigned to one of two conditions (*policy change* vs. *experiment*) in a between-subjects design.

The design of the study was nearly identical to that of Study 3a, with two changes. First, participants only evaluated the *negative/positive* stimuli (i.e., the left-most panel from Figure 2). Second, to make the survey fit on one page, we only included one of the three dependent variables ("How okay is it for the company to do this?") from Study 3a.

Results

Replicating our results from Study 3a, participants rated the experiments (M = 3.54) more favorably than their worst conditions (M = 2.41), t(239) = 7.06, $p < .001$.[7] These ratings are

[7] This analysis was done using a regression with fixed effects for each stimulus. We preregistered that we would also conduct a simple t-test collapsing across stimuli. The results are consistent, t(245) = 6.24, $p < .001$.

similar to MTurker ratings of identical stimuli in Study 3 (Experiments: M = 3.35; Worst Conditions: M = 2.26).[8]

Discussion

The results from Studies 3a and 3b are inconsistent with absolute experiment aversion, where people find all experimentation objectionable. Additionally, these results are inconsistent with a version of relative experiment aversion that is large enough to make an experiment less acceptable than its "worst" condition. In our next study, we apply the paradigm from Study 3 to directly examine the potential role of experiment aversion in the backlash to Kramer et al. (2014)'s Facebook experiment. Specifically, we assess whether the backlash may actually be attributed to the policies people were assigned to rather than experimentation per se.

STUDY 4: WAS FACEBOOK BACKLASH REALLY EXPERIMENT AVERSION?

As in Study 2, we investigated perceptions of an experiment based on Kramer et al. (2014). Unlike in Study 2, we used only stimuli that represented the specific conditions used in that experiment, rather than modifying certain aspects to find an "acceptable" version. We also used the same bipolar scales as Study 3, with labeled neutral midpoints, to evaluate policy changes and experiments.

Method

Sample. We recruited 748 participants on MTurk, of which 608 passed the attention check (41.3% female, M_{age} = 32.2 years). Participants were paid $0.30 for completing the study.

Design. The overall design of Study 4 was nearly identical to that of Study 3, but used different stimuli. Participants were randomly assigned to one of six conditions in a 2 (action:

[8] These numbers are not the same as those in Study 3a (and in the left panel of Figure 2) because in Study 3a we used a composite of three measures. Here, we compare only results for the question ("Is it okay for the company to do this?") that we used in both studies.

policy change vs. *experiment*) x 3 (policy combination: *sad/happy* vs. *no change/happy* vs. *sad/no change*) fully between-subjects design.

Participants in the *policy change* condition read that Facebook was considering making two policy changes (randomly selected from: sorting status updates to prioritize *happy* ones, to prioritize *sad* ones, or making *no change*). They then read that Facebook chose to implement one of the two policies. Participants in the *experiment* condition read that Facebook was considering running an experiment where they would randomly assign customers to two of the policy changes described above.

Measures. Participants answered the same acceptability questions from Study 3. However, because Facebook does not have an obvious competitor, we did not ask if participants would switch to a different company.[9] We average these two variables (r = .80) to construct the "acceptability index." Participants then indicated whether or not they had previously heard of Facebook taking similar actions in the past. This was collected to account for participants that may have been influenced by media coverage of the Facebook study.[10]

Results

Figure 3 shows the main results from Study 4. All three experiments (grey squares), even the experiment with ostensibly "good" conditions (i.e., *happy/no change*), were rated significantly below the acceptability midpoint (ts > 4.8, *ps* < .001). At first glance, this could be consistent with absolute or relative experiment aversion. However, this conclusion is not supported once we take into account the fact that the underlying policies are unacceptable even

[9] We exploratorily asked if participants would be inclined to cancel their Facebook membership; see pre-registration file.

[10] Most participants said that they had not heard of Facebook doing something similar (70.3% in the *experiment* condition and 82.2% in the *policy change* condition). Those with prior knowledge in the *experiment* condition rated Facebook's actions slightly more negatively (M = 2.77) than those with no prior knowledge (M = 3.06), t(301) = 1.75, *p* = .08. There was no difference between ratings in the *policy change* condition (*p* = .81). Therefore, we report results from all participants in our analysis.

outside of an experimental context. The lowest rated condition in each experiment was rated no higher than a 2.93 on a 7-point scale; significantly below the midpoint, ts > 9.4, *p*s < .001.[11]

As was the case in Study 3, when we directly compare the acceptability of experiments to the acceptability of their treatments' in the corresponding *policy change* conditions, we see that experimentation does *not* decrease the acceptability of the company's actions relative to some weighted average of its policy ratings. Indeed, experiments were again rated as at least marginally more acceptable than their worst conditions when considering each experiment individually, ts > 1.88, *p*s < .061, and significantly more acceptable when collapsing across all three experiments, $t(599) = 3.94$, $p < .001$.[12]

Discussion

Again, if there is relative experiment aversion, it is not large enough to push the experiment's ratings below the ratings of its policies. Thus, it is probable that participants were not reacting negatively to experimentation per se but to each experiment's underlying policies. Although the reaction to the Kramer et al. (2014) Facebook experiment is held up as evidence of a public distaste for corporate experiments, in Study 4 we find that Facebook probably did not face backlash because they ran an experiment, but because they implemented unacceptable policies. This suggests the public's reaction would have been even worse had Facebook modified how status updates are sorted for all (rather than for a random subset) of its users.

[11] The only specific policy that was rated above the midpoint was making no change (M = 5.10). Both sad status updates (M = 2.48), and happy status updates (M = 3.67) are viewed as unacceptable (all pairwise ts > 8.0, all ts vs. midpoint > 3.0).

[12] As indicated in our pre-registration, we ran a regression estimating ratings using fixed effects for each policy pair and an indicator for whether the participant rated a policy or an experiment. The coefficient for experiments was positive (b = .39; $p < .001$), indicating that experiments were rated more highly than policies when controlling for which policies participants saw.

INSERT FIGURE 3 ABOUT HERE

STUDY 5: RELATIVE EXPERIMENT AVERSION VS. CRITICAL CONDITION

Studies 3 and 4 demonstrate that relative experiment aversion, if it exists, may not be strong enough to drive ratings of an experiment below some weighted average of its policies. However, we cannot conclusively reject the existence of *some* relative experiment aversion. Even though the experiments were not rated worse than the least preferred policy, they were still rated below the equally-weighted average of its policies. This could be consistent with the critical condition account of experiment aversion if participants are taking a weighted average of their ratings of the two policies and giving more weight to the worse rated policy, as they might if they exhibit negativity bias (Skowronski & Carlston, 1989). However, this finding could also be consistent with the existence of moderate relative experiment aversion. For example, participants may be averaging their opinions of the policies and then applying some fixed "experiment penalty." Alternatively, participants' ratings of policies could be lower when those policies are part of an experiment. We ran Study 5 to more directly tease apart these two explanations by creating an experiment where both policies would be deemed equally acceptable. If there is relative experiment aversion, an experiment over both policies would be rated as lower than either, which would not happen if people evaluate experiments based on their critical conditions. We view this design as one which maximizes the ability to detect relative experiment aversion.

Method

Sample. We recruited 502 participants on MTurk, of which 406 passed the attention check (46.4% female, M_{age} = 35.0 years). Participants were paid $0.40 for completing the study.

Design. Participants were randomly assigned to one of two between-subjects conditions (*policy change* vs. *experiment*). We pretested the acceptability of 30 policies (see Supplement 4) and chose two that had nearly identical means (Ms = 5.54 and 5.59 out of 7) and distributions of responses (SDs = 1.40 and 1.32). The general design of Study 5 was similar to that of Studies 3 and 4. Participants read that a ride-sharing company (e.g., Uber, Lyft) was considering implementing two discounts (either a flat 10% discount or a $1 credit for every $10 spent) and either chose one of the two (*policy change* condition) or ran an experiment where they randomly assigned customers to receive one of the two discounts (*experiment* condition).

In both conditions, participants answered the following question: "How okay is it for the company to do this?" (1 = It's really bad; 4 = It's okay; 7 = It's really good).

Results

Participants rated both discounts (10% discount: M = 5.84; $1 credit for every $10 spent: M = 4.85) significantly above the midpoint, ts > 9.14, *p*s < .001, indicating that they viewed both discounts positively.[13] Participants rated the experiment that assigned participants to one of two discounts (M = 5.32) nearly identically to the *average* discount (M = 5.34), t(399) = .21, $p = .83$, and well above the least preferred discount (M = 4.61), t(399) = 5.24, $p < .001$. Participants in this study do not show even small levels of experiment aversion.[14]

[13] We should point out that the mean ratings of the individual discounts diverged more in Study 5 (Ms = 4.85 and 5.85) than they did in the pilot (Ms = 5.54 and 5.59). We believe that this is because evaluating only two discounts (compared to 10 in the pilot) made those discounts seem less similar.

[14] The 95% confidence interval for the difference between the acceptability of the experiment and the average policy is (-.21, +.26), thus we reject experiment aversion that is larger than .26 on our 7 point scale. With a pooled standard deviation of 1.19, we can reject experiment aversion having a Cohen's d > .22.

STUDY 6: ATTITUDES FROM ACTUAL CUSTOMERS

Our previous studies did not distinguish between reactions of customers and non-customers of the company running the experiment. Following suggestions of the review team, this study replicates our general design asking participants to evaluate experiments run and policies implemented by a company they regularly purchase from, Amazon. We also include a condition where the company is running an experiment with two negative policies to show that our results are robust when implementing objectionable policies may be unavoidable.

Method

Sample. We partnered with Lucid, a market research firm, to identify a nationally representative sample of regular Amazon customers (in our pre-registration, we defined *regular* users as those self-reporting making at least one purchase per month). Of the 3,681 people who started our survey, 2,185 successfully completed an attention check. Our final sample consists of the 1,304 regular Amazon customers among them (50.5% female; M_{age} = 44.2 years; 71.1% Amazon Prime members).

Design. Similar to Studies 3-5, participants were randomly assigned to one of six conditions, in a 2 (action: *policy change* vs. *experiment*) x 3 (policy combination: *negative/negative* vs. *negative/positive* vs. *positive/positive*) between-subjects design.

Participants in the *policy change* condition read that Amazon was considering two changes to its product recommendation system (presented in counterbalanced order). They evaluated how acceptable it would be if Amazon picked the first policy, then they were asked to evaluate how acceptable it would be if Amazon picked the second policy. Participants in the *experiment* condition read that Amazon was conducting an experiment where they randomly

assigned customers to one of the two changes to its product recommendation system and evaluated the acceptability of such an experiment.

In the *negative/negative* condition, the changes were (a) recommending the most profitable items or (b) recommending items that weren't selling well. In the *negative/positive* condition, the changes were (a) recommending the most profitable items or (b) recommending the most highly-rated items across the entire site. In the *positive/positive* condition, the changes were (a) recommending the most highly-rated items across the entire site or (b) recommending items that similar users have rated highly.

All participants were asked: "How okay is it for the company to do this?" (1 = It's really bad; 4 = It's okay; 7 = It's really good).

Results

Figure 4 shows that in this non-MTurk sample of actual (self-identified) Amazon customers evaluating experiments that would directly affect them, we replicate the "critical condition" finding—experiments are at least as acceptable as their worst condition is (all ts > 1.93, *p*s < .06). Again, there is no experiment aversion.[15]

INSERT FIGURE 4 ABOUT HERE

GENERAL DISCUSSION

Taken together, the results of our studies are inconsistent with *absolute* experiment aversion—experiments are considered acceptable if all policies tested in the experiment are

[15] As an exploratory analysis, we preregistered that we would compare results for customers with and without an Amazon Prime membership. Across the 9 evaluations (3 experiments and 6 policy changes), there were no statistically significant differences between self-reported Prime (N=927) vs non-Prime (N=377) customers. Among the 9 comparisons, *p*-values range from .09 to .50. See Supplement 5 for full set of results.

themselves acceptable. The results are also inconsistent with *relative* experiment aversion—experiments are considered to be at least as acceptable as their least acceptable policy. Experiments are not only acceptable under some circumstances, they are at least as acceptable as the worst policies they contain. We have called this the *critical condition* account of experiment evaluation.

These results are good news for companies that want to learn from experiments. Companies should not be more hesitant to run an experiment that includes a certain policy than they would be to implement that policy outright. A practical takeaway for organizations interested in running experiments is to first determine if their planned policy changes are objectionable (e.g., through a survey) and then run an experiment to determine which acceptable policy best achieves their desired objective. In these cases, companies are unlikely to face backlash for their experiments. Unfortunately, objectionable policies are sometimes unavoidable. Still, we find that experimentation with objectionable policies is preferred to implementing the worst policy by itself.

Limitations

We have identified three key limitations with our studies. The first limitation is that our samples consist primarily of people who volunteered to complete our studies, possibly excluding individuals who most strongly oppose data collection in general or experiments in particular. We are optimistic this is not a consequential limitation for two main reasons. First, our respondents did negatively evaluate experiments that included negative policies, indicating that they do not have universally positive opinions of experiments, and that they do discriminate between acceptable and unacceptable practices. Second, Study 3b surveyed a sample of non-academic

university staff, who do not regularly participate in experiments, and Study 6 used a non-MTurk sample provided by a market research firm. Their responses were indistinguishable from those of our MTurk samples. It is nevertheless impossible to obtain data on the attitudes of people who are unwilling to participate in an experiment.

The second limitation is that it is difficult to specify the threshold of acceptability that an action must reach to prevent a backlash. For example, a small group of motivated people (e.g., activists or media personalities) could be vocal enough to cause backlash against an experiment that most people find acceptable. At the same time, this concern applies to any action an organization can take and not solely experiments. Comparing the most extreme ratings across policy and experiment evaluations in our studies suggests experiments are not more polarizing than are policies. In Study 3a, for example, 12.5% of participants gave the negative policy the lowest possible rating and 7.6% of participants gave the experiment the lowest possible rating, a pattern that holds in all studies we run for which this comparison is possible.[16]

This also speaks to a larger issue of how different people may view different policy changes—what some may consider fine, others may find completely unacceptable. For this reason, we compared experiments to each participant's *least preferred* policy, rather than the average of each specific policy. Additionally, it is important to examine distributions of responses (beyond just means) to determine if a certain policy, although it may have a high mean, may be especially divisive (i.e., having a high variance). We encourage researchers and practitioners to pretest the acceptability of policies using surveys and measures like those we used in Studies 3 through 6.

[16] In Study 3b, 35.5% gave the lowest possible rating to the worst policy, compared to 9.8% for the experiment. In Study 4, these values are 20.7% and 12.9%, in Study 5, they are 2.5% and 1.0%, and in Study 6, they are 16.6% and 8.7%, respectively.

Finally, and perhaps most substantially, all of our scenarios are hypothetical. We simply cannot rule out the possibility that people will react differently to experiments that have actually occurred or that they were participants in than they would to a hypothetical study. For example, in some real world contexts, people could find a specific policy to be more objectionable when it is implemented as part of an experiment. We have not found any evidence that experimentation makes actions more objectionable, and we propose that experimentation generally does not make actions more objectionable. However, of course, we cannot unequivocally claim that an experiment will never make any policy more objectionable.

Experiment aversion is an interaction

Finally, there are many factors that could influence how acceptable experiments are. For example, much research has examined how people view the ethics of corporate practices that can be included in experiments, such collecting sensitive data (e.g., Awad & Krishnan, 2006; Culnan & Armstrong, 1999; Miyazaki, 2008), changing pricing practices (e.g., Bolton, Warlop, & Alba, 2003; Campbell, 1999; Haws & Bearden, 2006), or introducing new marketing strategies(e.g., Smith & Cooper-Martin, 1997).

Using the more specific context of our motivating example, it may be that Facebook's experiment was more objectionable because it involved emotions (or specifically *negative* emotions).[17] In addition, our review team proposed that perhaps people view an experiment as less acceptable when they participated in it, or when they are told about it after it has already been run. We report three studies that test these two hypotheses in the supplement (Studies S3-S5). We find that people prefer to hear about experiments before (rather than after) they are run (Study S3), that people's stated acceptability of an experiment is not affected by considering

[17] See Supplements 6-7 for descriptions of studies that test these questions.

having been a participant in it (Study S4), and that even when people consider having been assigned to the worst arm within an experiment, they rate the experiment overall as more acceptable than that worst arm (Study S5).

However, asking "Do these factors impact the acceptability of experiments?" will not teach us about experiment aversion, because these factors can be present in corporate actions both within and without an experiment. For example, a company can, outside of an experiment, take an action that affects consumers and inform them only after the fact. The critical question for the purposes of this paper, then, is "Do these factors impact the acceptability of experiments *more than they impact the acceptability of underlying policies*?" That is, is there an *interaction* between these factors and whether or not they are part of an experiment? In Studies S3 and S4, we find none of these hypothesized interactions (Study S3: $t(794) = .99$, $p = .32$; Study S4: $t(793) = .56$, $p = .58$). For example, in Study S3 we find that the negative effect of learning about an experiment after it is conducted (versus before it is conducted) is not larger than the negative effect of learning about a policy change after it is implemented (versus before it is implemented).

Of course, we did not test a completely exhaustive list of potential interactions. Similarly, we did not test any three-way interactions between these factors, so we cannot rule out those possibilities. For example, it is possible that people who are customers of a company show experiment aversion when they find out about an experiment after it is run, while people who are not customers of a company do not show experiment aversion regardless of when the experiment is disclosed. We propose that these interactions do not exist, or if they do exist, that they would not be large enough to be practically relevant. That said, we cannot rule out the possibility that any interaction does exist, and we encourage researchers to test for interactions. We expect all factors that influence opinion about experiments to also influence opinion about the acceptability

of policy changes. We propose that if something makes a policy unpopular, it will make an experiment that contains that policy unpopular, but not more so. Experiments are not unpopular, unpopular policies are unpopular.

Table 3 summarizes the contents of the online supplement:

INSERT TABLE 3 ABOUT HERE

References

Albergotti, R. (2014, July 1). Furor Erupts Over Facebook's Experiment on Users. *Wall Street Journal*. Retrieved from https://web.archive.org/web/20180105211315/https://www.wsj.com/articles/furor-erupts-over-facebook-experiment-on-users-1404085840

Awad, N. F., & Krishnan, M. S. (2006). The Personalization Privacy Paradox: An Empirical Evaluation of Information Transparency and the Willingness to Be Profiled Online for Personalization. *MIS Quarterly*, *30*(1), 13–28. https://doi.org/10.2307/25148715

Bolton, L. E., Warlop, L., & Alba, J. W. (2003). Consumer Perceptions of Price (Un)Fairness. *Journal of Consumer Research*, *29*(4), 474–491. https://doi.org/10.1086/346244

Campbell, M. C. (1999). Perceptions of Price Unfairness: Antecedents and Consequences. *Journal of Marketing Research*, *36*(2), 187–199. https://doi.org/10.1177/002224379903600204

Culnan, M. J., & Armstrong, P. K. (1999). Information Privacy Concerns, Procedural Fairness, and Impersonal Trust: An Empirical Investigation. *Organization Science*, *10*(1), 104–115. https://doi.org/10.1287/orsc.10.1.104

DellaVigna, S. (2009). Psychology and Economics: Evidence from the Field. *Journal of Economic Literature*, *47*(2), 315–372. https://doi.org/10.1257/jel.47.2.315

Folkes, V. S., & Kamins, M. A. (1999). Effects of Information About Firms' Ethical and Unethical Actions on Consumers' Attitudes. *Journal of Consumer Psychology*, *8*(3), 243–259. https://doi.org/10.1207/s15327663jcp0803_03

Gino, F. (2015, August 20). Companies Like Amazon Need to Run More Tests on Workplace Practices. *Harvard Business Review*. Retrieved from https://hbr.org/2015/08/companies-like-amazon-need-to-run-more-tests-on-workplace-practices

Goel, V. (2014, June 29). Facebook Tinkers With Users' Emotions in News Feed Experiment, Stirring Outcry. *The New York Times*. Retrieved from https://www.nytimes.com/2014/06/30/technology/facebook-tinkers-with-users-emotions-in-news-feed-experiment-stirring-outcry.html

Goldman, D. (2014, June 30). Facebook and Silicon Valley treat you like a lab rat. Retrieved February 19, 2019, from https://money.cnn.com/2014/06/30/technology/social/facebook-experiment/index.html

Greenfield, R. (2014, July 28). OkCupid's Human Experiments Are Way Creepier Than Facebook's. Retrieved February 19, 2019, from https://www.fastcompany.com/3033645/okcupids-human-experiments-are-way-creepier-than-facebooks

Haws, K. L., & Bearden, W. O. (2006). Dynamic Pricing and Consumer Fairness Perceptions. *Journal of Consumer Research*, *33*(3), 304–311. https://doi.org/10.1086/508435

Hill, K. (2014, July 28). OkCupid Lied To Users About Their Compatibility As An Experiment. Retrieved February 19, 2019, from https://www.forbes.com/sites/kashmirhill/2014/07/28/okcupid-experiment-compatibility-deception/

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790. https://doi.org/10.1073/pnas.1320040111

Luca, M. (2014, July 29). Were OkCupid's and Facebook's Experiments Unethical? *Harvard Business Review*. Retrieved from https://hbr.org/2014/07/were-okcupids-and-facebooks-experiments-unethical

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*, 273–332.

Meyer, R. (2014, June 28). Everything We Know About Facebook's Secret Mood Manipulation Experiment. Retrieved February 19, 2019, from https://www.theatlantic.com/technology/archive/2014/06/everything-we-know-about-facebooks-secret-mood-manipulation-experiment/373648/

Miller, G., & Mobarak, A. M. (2014). Learning About New Technologies Through Social Networks: Experimental Evidence on Nontraditional Stoves in Bangladesh. *Marketing Science*, *34*(4), 480–499. https://doi.org/10.1287/mksc.2014.0845

Miyazaki, A. D. (2008). Online Privacy and the Disclosure of Cookie Use: Effects on Consumer Trust and Anticipated Patronage. *Journal of Public Policy & Marketing*, *27*(1), 19–33. https://doi.org/10.1509/jppm.27.1.19

Montaguti, E., Neslin, S. A., & Valentini, S. (2015). Can Marketing Campaigns Induce Multichannel Buying and More Profitable Customers? A Field Experiment. *Marketing Science*, *35*(2), 201–217. https://doi.org/10.1287/mksc.2015.0923

Muse, T. (2014, August 4). The Facebook Experiment: What It Means For You. Retrieved February 18, 2019, from https://www.forbes.com/sites/dailymuse/2014/08/04/the-facebook-experiment-what-it-means-for-you/

Rozin, P., & Royzman, E. B. (2001). Negativity Bias, Negativity Dominance, and Contagion. *Personality and Social Psychology Review*, *5*(4), 296–320. https://doi.org/10.1207/S15327957PSPR0504_2

Rudder, C. (2014). We experiment on human beings! Retrieved from https://web.archive.org/web/20170316051630/https://theblog.okcupid.com/we-experiment-on-human-beings-5dd9fe280cd5

Schroepfer, M. (2014, October 2). Research at Facebook | Facebook Newsroom. Retrieved February 19, 2019, from https://newsroom.fb.com/news/2014/10/research-at-facebook/

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin*, *105*(1), 131–142. https://doi.org/10.1037/0033-2909.105.1.131

Smith, N. C., & Cooper-Martin, E. (1997). Ethics and Target Marketing: The Role of Product Harm and Consumer Vulnerability. *Journal of Marketing*, *61*(3), 1–20. https://doi.org/10.1177/002224299706100301

Stampler, L. (2014, July 28). Facebook Isn't the Only Website Running Experiments on Human Beings. Retrieved February 19, 2019, from http://time.com/3047603/okcupid-oktrends-experiments/

Sudhir, K., Roy, S., & Cherian, M. (2016). Do Sympathy Biases Induce Charitable Giving? The Effects of Advertising Content. *Marketing Science*, *35*(6), 849–869. https://doi.org/10.1287/mksc.2016.0989

Wang, Y., Lewis, M., Cryder, C., & Sprigg, J. (2016). Enduring Effects of Goal Achievement and Failure Within Customer Loyalty Programs: A Large-Scale Field Experiment. *Marketing Science*, *35*(4), 565–575. https://doi.org/10.1287/mksc.2015.0966

Zoumpoulis, S., Simester, D., & Evgeniou, T. (2015, November 12). Run Field Experiments to Make Sense of Your Big Data. *Harvard Business Review*. Retrieved from https://hbr.org/2015/11/run-field-experiments-to-make-sense-of-your-big-data

Table 1. Stimuli and measures for Study 1

Policy changes			
Context	**Bad**	**Good**	**Very good**
1. Shipping	Slower delivery	Faster delivery	Much faster delivery
2. Company gym	$5 penalty for not going	$5 bonus for going	$10 bonus for going
3. Product recommendations	Poorly rated products	Highly rated products	Highest rated overall

Measures of Acceptability
Participants indicated agreement (1=Strongly Disagree; 7=Strongly Agree), with these statements.
Acceptability of policy changes
1. It is okay for the company to do this.
2. If I were [an employee/a customer], I would object to this. (reverse-coded)
3. If I were [an employee/a customer] and was asked, I would agree to this.
Acceptability of experiment
4. It is immoral to run this experiment (reverse-coded)
5. People in this experiment are being treated like guinea pigs (reverse-coded)
6. The company should be not allowed to run this experiment (reverse-coded)

Notes: Participants in the policy change condition rated all nine policy changes. Participants in the experiment conditions rated one of nine experiments created by pairing two policy changes within a context. The pairs consisted of bad/good, control/good or good/very good. Control consists of keeping the status quo (e.g., shipping item as promised). The average of questions 1-3 is the policy acceptability index, the average of questions 4-6 the experiment acceptability index.

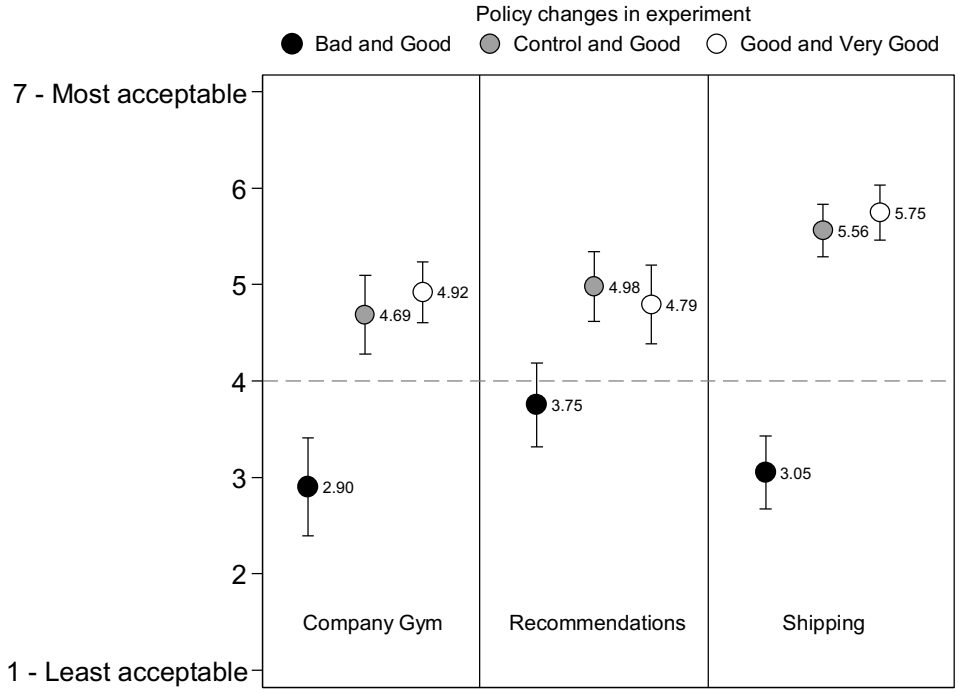
Table 2. Overview of study design and contributions

Study 1	• Test of absolute experiment aversion. • People find experiments with unambiguous benefits acceptable.
Studies 2a & 2b	• Extends Study 1 with more realistic stimuli. • Acceptability of conditions predicts acceptability of experiments.
Study 3a	• Direct comparison of experiments with underlying conditions. • Experiments rated at least as acceptable as worst condition is. • Results hold for variety of stimuli.
Study 3b	• Replicates Study 3a results using a sample that does not regularly volunteer for experiments.
Study 4	• Best known example of experiment aversion is not an instance of experiment aversion.
Study 5	• Experiments with similar and positively-viewed policies are rated identically to the average policy.
Study 6	• Replicates Study 3 using company's own customers. • Adds condition with experiments using two negative policies.

Table 3. Index of supplementary materials (available from http://osf.io/z39aq)

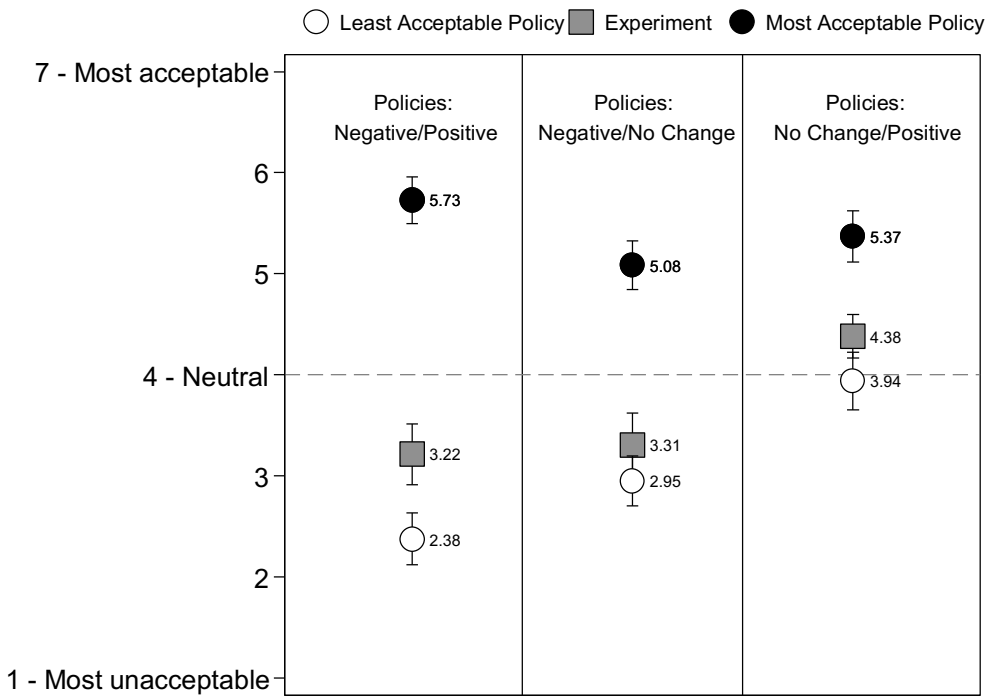
Section	Pages
Supplement 1. Additional Study 1 analysis	2-3
Supplement 2. Full list of Study 3 stimuli	4
Supplement 3. Additional analyses for Study 4 included in pre-registration	5-7
Supplement 4. Study 5 pilot results	8-9
Supplement 5. Prime vs. Non-Prime customers in Study 6	10
Supplement 6. Overview of studies not included in main manuscript	11-13
Supplement 7. Details and results for studies not included in main manuscript	14-25

Figure 1. Experiments without bad policies (gray and white circles) are rated positively (Study 1)



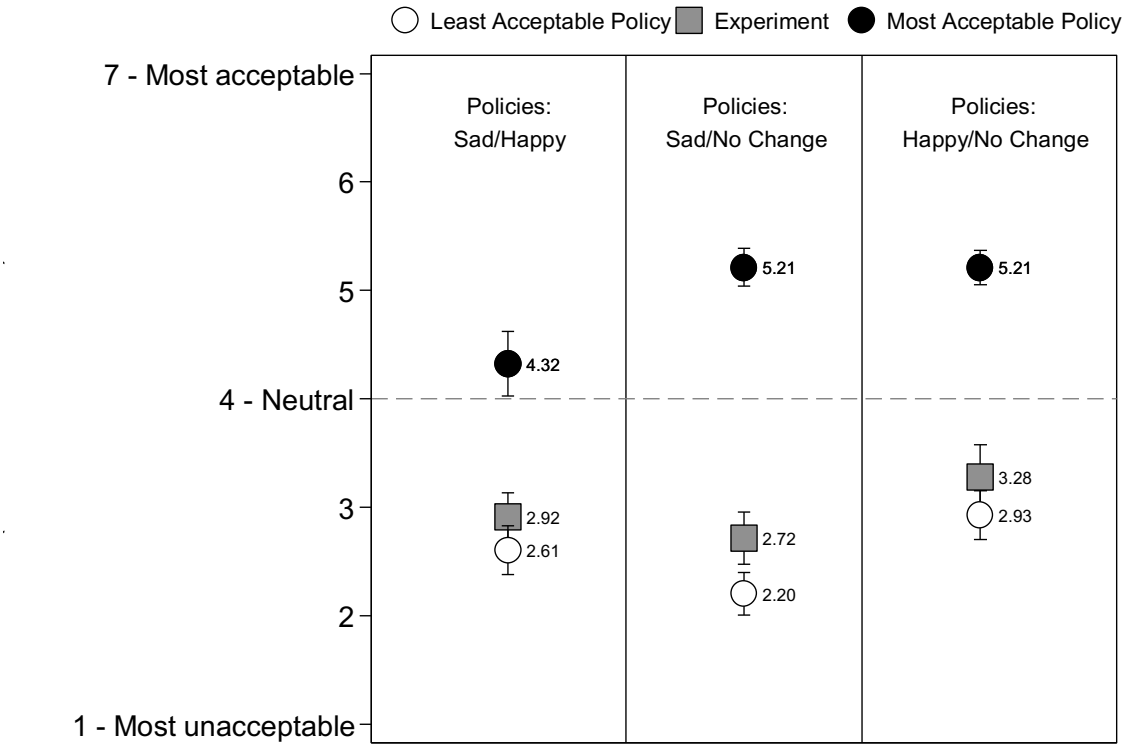
Notes: Each participant (N = 452) rated the acceptability of one experiment (out of 9 possible experiments). Markers depict sample averages; error bars represent 95% confidence intervals.

Figure 2. Experiments (gray squares) are no less acceptable than their least acceptable condition (white circles) (Study 3a)



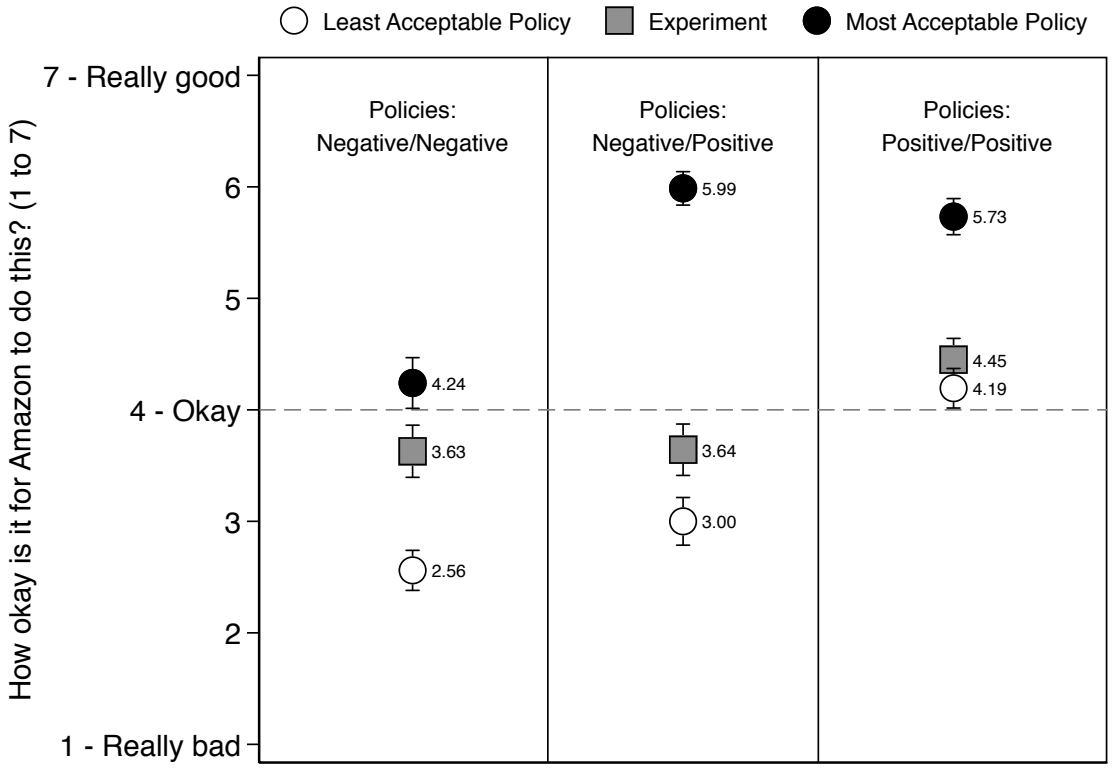
Notes: Each participant (N=423) rated the acceptability of a company choosing one of two policies or running an experiment using those two policies as conditions. The policies involved a negative change, a positive change, or no-change. Circular markers depict means evaluation of each policy, squared markers the evaluations of the experiment that combines them. Error bars represent 95% confidence intervals.

Figure 3. Facebook experiments (gray squares) are no less acceptable than their least acceptable condition (white circles) (Study 4)



Notes: Each participant (N = 601) rated the acceptability of Facebook changing how status updates are sorted or of running an experiment randomly assigning users to one of those changes. Circular markers depict mean evaluations of the least and most acceptable change in the pair, squared markers depict mean evaluations of an experiment randomly assigning users to them. For example, the first panel shows that people evaluating sorting status updates by sad/happy rated the worst of these with M=2.70, the highest with M = 4.22, and an experiment with M = 2.92. Error bars represent 95% confidence intervals.

Figure 4. Amazon customers rate potential Amazon experiments (gray squares) more positively than their worst policies (white circles) (Study 6)



Notes: Each participant (N = 1,304) rated the acceptability of Amazon changing how they recommend products to customers or running an experiment randomly assigning customers to one of those changes. Circular markers depict mean evaluations of the least and most acceptable change in the pair, squared markers depict mean evaluations of an experiment randomly assigning users to them. Error bars represent 95% confidence intervals.