Psychological Methods

Harvesting Heterogeneity: Selective Expertise Versus Machine Learning

Rumen Iliev, Alex Filipowicz, Francine Chen, Nikos Arechiga, Scott Carter, Emily Sumner, Totte Harinen, Kate Sieck, Kent Lyons, and Charlene Wu

Online First Publication, October 7, 2024. https://dx.doi.org/10.1037/met0000640

AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Methods

ISSN: 1082-989X

https://doi.org/10.1037/met0000640

Harvesting Heterogeneity: Selective Expertise Versus Machine Learning

Rumen Iliev[1], Alex Filipowicz[1], Francine Chen[1], Nikos Arechiga[1], Scott Carter[1], Emily Sumner[1], Totte Harinen[1, 2], Kate Sieck[1], Kent Lyons[1, 3], and Charlene Wu[1]

[1] Toyota Research Institute, Los Altos, California, United States

[2] Airbnb, San Francisco, California, United States

[3] Inovo Studio, Los Altos, California, United States

Abstract

The heterogeneity of outcomes in behavioral research has long been perceived as a challenge for the validity of various theoretical models. More recently, however, researchers have started perceiving heterogeneity as something that needs to be not only acknowledged but also actively addressed, particularly in applied research. A serious challenge, however, is that classical psychological methods are not well suited for making practical recommendations when heterogeneous outcomes are expected. In this article, we argue that heterogeneity requires a separation between basic and applied behavioral methods, and between different types of behavioral expertise. We propose a novel framework for evaluating behavioral expertise and suggest that selective expertise can easily be automated via various machine learning methods. We illustrate the value of our framework via an empirical study of the preferences towards battery electric vehicles. Our results suggest that a basic multiarm bandit algorithm vastly outperforms human expertise in selecting the best interventions.

Translational Abstract

Over the last century, behavioral science has made tremendous progress in describing, predicting, and understanding human behavior, yet it still faces major challenges when it is applied to concrete real-world problems. Many of these challenges are associated with the great amount of variability between groups, individuals, stimuli, and contexts. A common approach to address this problem among academic researchers has been to focus on generalizability by using more representative samples and more comprehensive theoretical models. Here, we propose an alternative approach, which is based on separating the role of behavioral expertise in basic versus applied research. We decompose the role of expertise in applied and basic research, and we suggest that selective expertise in applied research can be optimized by using machine learning approaches. We illustrate our perspective by comparing the cost and benefits of a classical experimental design and a basic machine learning optimization algorithm. Our results suggest that in many tasks where behavioral data are available in real time, machine learning will be playing an increasingly important role in selecting and administering behavioral interventions.

Keywords: expertise, heterogeneity, multiarm bandits, applied research, environmental psychology

Supplemental materials: https://doi.org/10.1037/met0000640.supp

It has recently been suggested that behavioral science will not be able to meet the growing need for applied behavioral research without a "heterogeneity revolution." More specifically, Bryan et al. (2021) argued that part of this gap stems from academic behavioral science's focus on main behavioral effects, which are often observed in lab-based studies, but these studies largely ignore the impact contextual factors and individual differences play in the real world. We agree with this sentiment, yet we argue that such revolution also requires a greater differentiation between applied and theoretical behavioral methods. We specifically propose that in the context of

Rumen Iliev https://orcid.org/0000-0002-9619-4166

Rumen Iliev served as lead for conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization, writing–original draft, and writing–review and editing. Alex Filipowicz contributed equally to formal analysis and methodology. Kate Sieck served in a supporting role for supervision. Kent Lyons served in a supporting role for supervision. Charlene Wu served as lead for project administration and supervision. Alex Filipowicz, Nikos Arechiga, and Totte Harinen contributed equally to conceptualization. Francine Chen, Nikos Arechiga, Scott Carter, and Emily Sumner contributed equally to investigation. Alex Filipowicz, Francine Chen, Nikos Arechiga, Scott Carter, Emily Sumner, Totte Harinen, Kate Sieck, and Charlene Wu contributed equally to writing–original draft.

Correspondence concerning this article should be addressed to Rumen Iliev, Toyota Research Institute, 4440 El Camino Real, Los Altos, CA 94022, United States. Email: rumen.iliev@tri.global

applied research, some aspects of behavioral expertise can be enhanced or entirely replaced by machine learning (ML) algorithms. We illustrate this point with a comparison between classical experimental design (CED) and a basic ML algorithm for selecting the best intervention[1] and demonstrate the advantages of automating the selection process.

From Crisis of Confidence to Acceptance of Heterogeneity

For observers following trends in behavioral science over the last decade, there must be an apparent discrepancy between the critical self-evaluation coming from behavioral scientists and the growing appetite for applied behavioral research coming from nonacademic institutions and organizations. On the one hand, the validity of core theoretical and empirical behavioral science findings have recently come under serious scrutiny. In fact, a common epithet attached to modern behavioral science is "crisis": crisis in confidence (Earp & Trafimow, 2015), credibility (Horgan, 2016), replicability (Maxwell et al., 2015), and generalizability (Yarkoni, 2020). On the other hand, behavioral science is becoming an increasingly popular method for solving problems in the real world. Behavior is easier to observe and study than ever before (Harlow & Oswald, 2016; King, 2011) and there has been a growing appreciation of the importance of behavioral research when aiming at understanding or influencing human behavior at scale. Governments, nongovernmental organizations, and industries are actively testing and applying insights from behavioral research, leading to tangible outcomes (Halpern, 2015; Internal Revenue Service, 2017; Van Bavel et al., 2020; Walton & Wilson, 2018; The White House, 2015).

How can we reconcile this apparent contradiction? A recently emerging consensus is that although behavioral interventions can be very effective, there is high a priori uncertainty about how effective any particular intervention will be (Milkman et al., 2021). The term often associated with this uncertainty is heterogeneity of outcomes. The term heterogeneity in psychological literature is used in three contexts. One is associated with moderation effects, where substantial group differences for the same treatment are observed (Bryan et al., 2021). A second meaning is about large differences in effect sizes of conceptually related interventions (Linden & Hönekopp, 2021). A third possible meaning of the term heterogeneity includes individual differences (Molenaar & Campbell, 2009) which are not necessarily associated with a particular group-level moderators, as in the first meaning, but stem from longitudinal differences in individual histories or idiosyncratic contexts.[2] In this article, when we refer to heterogeneity of outcomes, we include all three meanings of the term, and we differentiate between them only when required by the argument.

Multiple examples of successful lab demonstrations and real-world implementations are accompanied by multiple cases of failed replications. A vivid illustration is loss aversion, which is one of the most broadly known results coming from behavioral research. The initial claim was the general principle that "losses loom larger than gains" (Kahneman & Tversky, 1979), a finding that was originally reinforced by a broad range of lab and real-world studies. However, more recent subsequent research has found multiple factors that influence the strength, presence, or even the directionality of the effect. Loss aversion turns out to depend on culture (Wang et al., 2017), gender (Bouchouicha et al., 2019; Rau, 2014; Wang et al., 2017), wealth (Xie et al., 2018), and the amounts at stake (Harinck et al., 2007) among other factors. After four decades of empirical tests, the existence of a general loss aversion is a hotly debated topic (Gal & Rucker, 2018; Simonson & Kivetz, 2018), accompanied by an emerging consensus that there is a large degree of heterogeneity associated with the effect (Gächter et al., 2022; Sproul & Michaud, 2017).

Although loss aversion is an extreme example in terms of its impact and popularity, its history is fairly representative of the general trend toward diminishing main effects and increased heterogeneity in follow-up studies. In a seminal paper, Henrich et al. (2010) showed that the effect sizes of classical psychological studies decrease or disappear when the study populations are different from the typical western college students sample. More recently, Linden and Hönekopp (2021) reviewed 150 meta-analyses and found that there is a very high degree of heterogeneity of results, with "powerful moderators conspicuously absent." The authors suggested that unexplained heterogeneity is an indicator of a poor state of a psychological theory, and that its reduction should be an explicit research goal. Furthermore, Bryan et al. (2021) suggested that acknowledging and properly addressing heterogeneity is the only way behavioral research can help solve real-world problems. They additionally suggest that psychological studies should more readily rely on heterogeneous and generalizable pools of participants and that analyses should be explicitly focused on detecting moderation effects.

The heterogeneity of outcomes provides a major challenge to the value of expertise in behavioral science. Designing successful interventions requires some anticipation of the intervention's potential results, yet results are difficult to predict a priori, especially when existing findings generalize poorly across contexts and populations. Although researchers have been aware of these difficulties (Yarkoni, 2020), we have only recently begun estimating the discrepancy between expert predictions and real world results. For example, DellaVigna and Linos (2022) compared the difference in effect sizes between behavioral interventions conducted in laboratory settings and interventions conducted in the real world. On average, they found that while lab-based interventions increased desired behavior by 8.7% on average, their effectiveness falls to a 1.4% increase when conducted in the real world. As an additional example, Milkman et al. (2021) ran a large-scale study testing the effect of 54 interventions designed to encourage physical exercise. As a part of their experiment, the authors also asked lay people, university professors, and applied behavioral science practitioners to predict the effectiveness of these different interventions. Rather surprisingly, the correlation between the predicted and observed efficacy of the different interventions was virtually zero, and predictions from experts were no more accurate than predictions from lay people. Such findings prompt reconsidering the role of expertise in the context of applied research.

It is important to note that the problem posed by the heterogeneity of outcomes is not only about potentially failing to find a beneficial intervention, but also the risk of launching a potentially harmful one.

[1] Sometimes behavioral scientists distinguish between interventions and experimental manipulations, based on the particular experimental design, duration, or cost. Here, however, we use only the term intervention, where intervention broadly means a change which is under the control of the researcher and is aimed at achieving a particular outcome via some causal path (Pearl, 2010; Pearl & Mackenzie, 2018).

[2] We thank to Alexander Christensen for pointing at this third aspect of heterogeneity.

A recent example comes from the work of Goldberg et al. (2021), who used targeted advertising to influence Republicans' view on climate change. The authors observed the effect that they had hypothesized, namely, that a well-designed advertising campaign can shift the recipient's attitudes around climate change and its causes. However, in a subsequent analysis, Harinen et al. (2021) used modern ML algorithms to analyze heterogeneity and observed substantial group differences. Their analysis found that although Republicans' views around climate change were on-average influenced by the authors' intervention, there were some subgroups for which the intervention backfired. More specifically, the treatment was most effective among young, non-White Republicans, but had the opposite effect among middle-aged, White Republicans, leading to a decrease in the belief that climate change was happening and that it was caused by human activity (Harinen et al., 2021). Relying on CED to test the efficacy of interventions provides only a very rudimentary test, and not only leads to missed opportunities for developing more powerful interventions, but can mask backfire effects that may be difficult or even impossible to predict based on expertise alone.

Expertise in Basic Versus Applied Research

In this section, we propose a basic decomposition of behavioral expertise and assess how different types of expertise might play different roles in applied versus basic research. For our current purposes, we can decompose a basic empirical research project into five steps (see Figure 1): (a) We start with a research question. (b) We generate various hypotheses that are intended to validate or falsify some theoretical inferences. We consider different methods to test those hypotheses, and we also consider various operationalizations of the relevant variables. (c) We next focus on one or a few hypotheses, and we choose a particular set of methods and operationalizations to test these hypotheses. (d) We run the required empirical tests. (e) Depending on the results, we provide an answer to the research question we started with, and, typically, update the existing theoretical knowledge. We can also divide a typical applied research process in the same five steps, except that we replace a research question with a research goal, hypotheses with interventions or treatments, and instead of theory update, the final step is launching the selected intervention at scale.

Although this five-step process is a somewhat crude summary of the life-cycle of a generic research project, it is able to capture at least some of the depth of behavioral expertise researchers rely on. The first type is generative expertise, which refers to the creative part of the human expert and allows us to generate hypotheses, methods, and operationalizations. The second type is selective expertise, which is related to our ability to choose which hypotheses to test and what research methods to use in order to test them. Last, for each step of the research process, there is methodological expertise, which comes from education and/or from practice, and includes technical knowledge related to the design and implementation of experiments, data analysis, and making logical or statistical inferences based on results. For example, part of a behavioral scientists' methodological expertise can include knowledge about placebo effects, Hawthorne effects, spillover effects, carryover effects, Rosenthal effects, demand characteristics, reliability, validity, and generalizability.

Although these three types of expertise are relevant for both for basic and for applied research, there is an important difference that is central to this article. In basic research, all three types of expertise are very hard to quantify. Questions such as "Would the research project have been more successful if a different set of hypotheses were generated at Step 2," or "What if the project tested a different population at Step 4?" rarely have objective answers because they involve multiple qualitative components. Arguments over the appropriateness of those components can easily result in a substantial disagreement between experts. Similarly, in applied research, methodological and generative expertise are also difficult to quantify or to measure objectively. However, this is not the case with selective expertise. Unlike basic research, in applied research the selective expertise can be objectively measured as the accuracy of a priori predictions. If an expert is asked to select a number of interventions at Step 3 based on the likelihood that each intervention will have the desired result, their predictions can be objectively compared to the results from all generated interventions. The correspondence between the predicted and the observed relative effectiveness of the selected intervention can be used as a metric of selective expertise. Furthermore, because in the applied case, we have objective answers, we might also be able to enhance or fully automate the process of selection at Step 3. We will next discuss how this automation could be undertaken.

Boosting Selective Expertise via ML

Both basic and applied behavioral research rely heavily on CED, where an intervention is hypothesized to lead to a particular effect, and then an empirical test with a predetermined sample size is conducted to verify the research hypothesis. In contrast to basic research, however, in applied research, we are not interested in all results, but only in choosing the most effective intervention. Since applied settings prioritize results over understanding, ML methods can greatly accelerate the time and cost required to find an effective intervention.

If each participant is assigned to a single intervention,[3] the problem of empirically finding the best intervention is known as a multiarmed bandit (MAB). The name comes from an analogy with finding the best paying slot machine in a casino. Currently, this is a popular research topic and there are multiple algorithms for solving this problem.[4] The most simple approach, known as ϵ-greedy, is to start testing various alternatives, choosing the currently best alternative with a probability ϵ or any other alternative with probability $1 - \epsilon$. While there are better algorithms available,[5] for this article, we will use only the ϵ-greedy algorithm.

It is worth noting that while MAB algorithms have been known for a long time (Thompson, 1933), they have not been popular among researchers before the "online behavior" era since they rely on real-time data processing and fast feedback. Currently, tech-industries are spearheading the application of those methods at scale (Amadio, 2020; Amat et al., 2018), with social and behavioral scientists also becoming increasingly interested in applying them to academic projects (Jahanbakhsh, 2020; Offer-Westort et al., 2021; Schwartz et al., 2017). However, to the best of our knowledge, there is no systematic work that evaluates the role of expertise in the context of experimentation systems that employ MAB

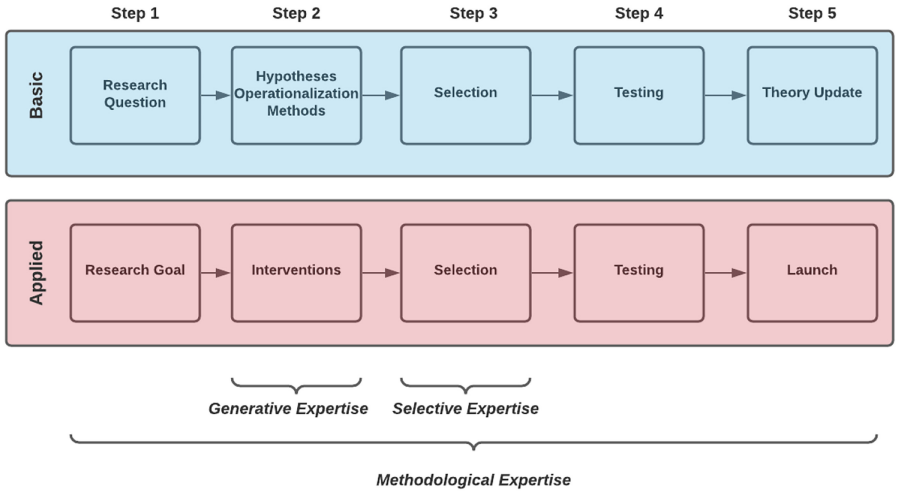
[3] In this article, we focus on the single intervention case, which is the most simple example and easy to illustrate, but ML methods could similarly be applied to more complex designs, such as multiple simultaneous interventions or sequences of interventions.

[4] For a classical introduction to the topic see Sutton and Barto (2018).

[5] State-of-the-art contextual MAB algorithms can be found at https://vowpalwabbit.org.

Figure 1

A Five-Step Process for Basic and Applied Experiment-Based Empirical Research



Note. Generative expertise is mainly relevant to Step 2, selective expertise is mainly relevant to Step 3, and methodological expertise is relevant to all five steps.

algorithms. In the remainder of the article, we raise five research questions focused on the potential advantages of MAB over CED, and we run a series of studies to provide answers to those questions.

Research Questions

1. Is the amount of heterogeneity of outcomes sufficient to justify the application of ML methods?
2. Can heterogeneity be predicted a priori?
3. Are experts better at predicting heterogeneity than nonexperts?
4. How much advantage does a MAB have over CED in terms of sample size and/or time-cost?
5. How much does expertise help a MAB algorithm?

Study 1: The Amount of Heterogeneity

The high-level goal of this project is to explore the advantages of ML algorithms for selecting interventions. We embody this goal in a concrete, practical problem specific to the workaday needs of our group. Namely, we are investigating ways to determine the most efficient messages for increasing the preferences for battery electric vehicles (BEVs) over internal-combustion engine vehicles (ICEVs). It is broadly accepted that vehicle electrification will play a major role in the future efforts to reach carbon neutrality (Burnham et al., 2021; Jenn, 2020), yet various studies have found that a substantial proportion of drivers are still reluctant to switch from an ICEV to a BEV (Brückmann et al., 2021). Although governments across the globe are planning to rapidly shift from ICEVs to BEVs (Shepardson & Klayman, 2021), there are indicators that drivers' preferences and attitudes are still lagging behind the oncoming technological and market change. For example, in a recent representative-sample survey in the United States, only 31% of the participants indicated that they are considering a BEV as their next car (Consumer Reports, 2020). While most of the efforts associated with the shift toward BEVs will be in terms of building infrastructure and developing new technologies, to meet the ambitious carbon neutrality goals set by governments and businesses some of the efforts will need to be focused on changing consumer preferences. In this study, therefore, we aim at finding the most persuasive short messages for changing people's preference toward BEVs.

Method

Participants

A total of 4,136 participants were recruited via Amazon Mechanical Turk (50% male, 47% female, 3% other, or missing). The participants were paid $3. The study followed Institutional Review Board (IRB) guidelines and regulations and was approved by Western IRB. The participants gave their informed consent at the beginning of the study.

Stimuli

When designing preference change interventions, a common choice is between affective- or cognitive-based interventions (Zajonc & Markus, 1982). In this project, we have focused on cognitive interventions alone, yet the overall framework we propose here is suitable for both types of interventions. For this study, a cognitive intervention is a short snippet of information that emphasizes a particular objective advantage of BEVs over ICEVs. Since the potential difference in the effectiveness of the interventions is of central interest, we followed a rigorous process when creating them. First, we ran a literature review on the barriers and motivators associated with BEVs. We found three common topics: range/charging, cost, and environmental impact. In addition, we ran our own exploratory survey (not described in this article), and found that people were also concerned with reliability and utility of BEVs, and a broader set of topics that we labeled "other." Next, we took the resulting six topics, and searched the web

for various pieces of pro-BEV information related to those topics. Our search resulted in 35 short statements, such as

> 80% of BEV charging happens at home, and most trips do not involve public charging.

The full intervention list can be found in the online supplemental materials.

Procedure

Each participant answered a series of demographic questions, followed by a question about their initial BEV preferences. The preferences were measured on a 0–100 Likert scale, where greater numbers meant stronger preference for BEVs. After the initial preference measure, the participants were randomly presented one of the 35 interventions, and were then asked to indicate their preference again. This sequence of random[6] intervention followed by preference measure was repeated five times, but for the purposes of the current article, we are analyzing only the preference shift due to the first intervention.[7] The procedure took 15 min on average.

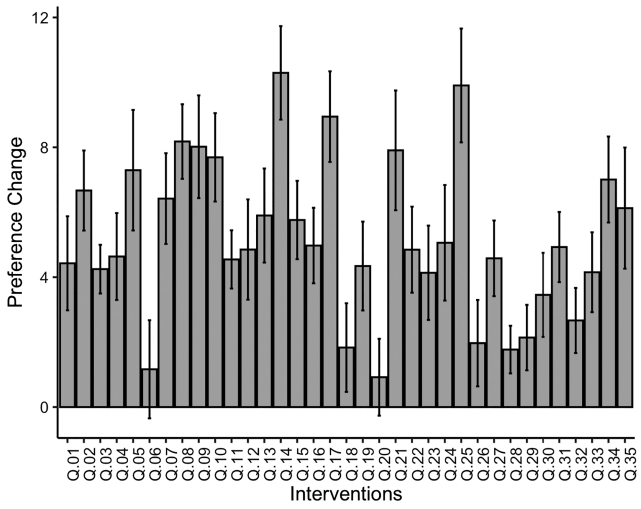
Results and Discussion

Preference shift was computed as the difference between the post-intervention and preintervention preferences for BEVs, where positive numbers indicate the greater preference for BEVs after the intervention. The average shift across all interventions was 5.19, $SD = 2.42$, intervention-level range = 0.92–10.29. The effects for each of the 35 interventions are presented in Figure 2.

Next, we looked at the degree to which the heterogeneity due to different interventions was consistent across demographic groups. We used a basic demographic split based on dichotomized age (younger vs. older than 35) and gender. The average preference shifts for each intervention and for the four demographic groups based on gender and age are presented in Figure 3. When taking both interventions and demographics into account, the intervention-level range of the effect sizes was from −4.29 to 20.13. Importantly, the range of effects we observe when splitting by age and gender are larger than when examining the main effects alone. This critically indicates that in the context of our experiment, adding even basic demographic information will be highly beneficial in achieving a maximal effect size.

Figure 2

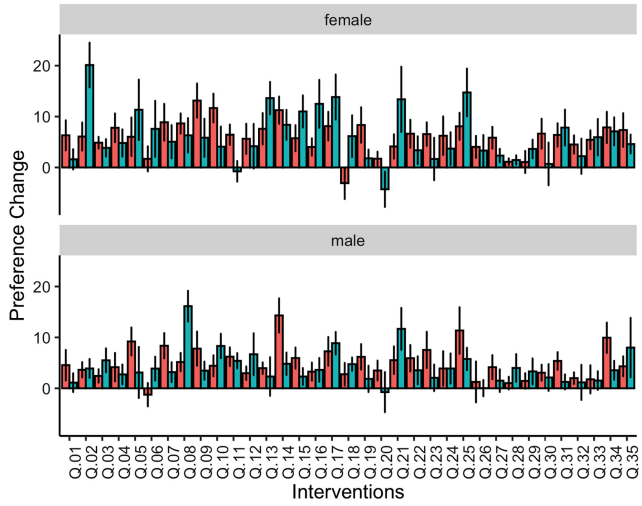
Preference Shift for Each of the 35 Interventions



Note. Positive numbers indicate an increase in the preference toward BEVs after the intervention. The error bars indicate ±1 *SE*. BEVs = battery electric vehicles; *SE* = standard error.

Figure 3

Preference Shift by Intervention and Demographics (Age and Gender)



Note. Positive numbers indicate an increase in the preference toward BEVs after the intervention for younger (green) and older (red) participants. The error bars indicate ±1 *SE*. BEVs = battery electric vehicles; *SE* = standard error.

In addition to considering range, we also examined the correspondence of effect sizes across demographics. Did the same interventions lead to similar preference change for different age and gender groups? Using Pearson's product–moment correlation, we found a moderate correspondence between the intervention effectiveness for older males and older females, but a rather low correspondence across ages, or between younger males and younger females. For exact correlations see Table 1.

Study 2: Evaluating Selective Expertise

The high heterogeneity observed in Study 1 would be a concern only if it cannot be predicted a priori. In this study, we test to what degree the empirically observed heterogeneity can be guessed in advance.

Method

Participants

We placed an ad in an internal communication channel asking for help with this study. Fifteen subjects volunteered to participate. They were familiar with the project, but were not aware about the results of

[6] Interventions were chosen randomly without replacement.

[7] Initially, we were aiming for a fully fledged reinforcement learning platform which was going to interactively choose sequence of interventions, but we had to adjust the original goal to a first intervention MAB algorithm only.

Table 1
Correlations Between the Intervention Effectiveness for Different Demographic Groups

Demographic groups	older_male	younger_female	younger_male
older_female	.63	.15	.21
older_male		.27	.15
younger_female			.30

the interventions. The participants came from various professional backgrounds, including ML, data science, human–computer interaction, and behavioral science.

Stimuli and Procedure

The participants were asked if each of the 35 interventions from Study 1 would increase or decrease the overall preferences for BEVs. For each intervention, the participants answered on a 7-point Likert scale, ranging from −3 (*it will definitely decrease the preference*) to 3 (*it will definitely increase the preference*).

Results and Discussion

First, we collapsed across participants and computed an overall score of the predicted effect sizes for each of the 35 interventions. We correlated these predictions with the observed effects in Study 1, finding a moderate agreement ($r[34] = .32$). The accuracy of the individual predictions ranged from −.05 to .37. Next, we checked if self-reported behavioral expertise was associated with more accurate predictions. We correlated individual predictive accuracy with expertise and we found that expertise was associated with higher accuracy ($r[14] = .57$, $p = .03$). We also split the participants into a high behavioral expertise group ($n = 5$)[8] and a low behavioral expertise group ($n = 10$) and found that the overall accuracy of the high expertise group was numerically higher ($r[34] = .37$) than the accuracy of the low expertise group ($r[34] = .24$), but the difference was not statistically significant.

We also ran an additional analysis, using a mixed-effects model (Kuznetsova et al., 2017) predicting human judges expectations from the empirically observed changes in preference. In a first model, we entered only empirically observed effect sizes as fixed effect, and we allowed random intercepts for individual judges and for individual interventions. The empirically observed changes were marginally associated with judges' expectations ($t[33] = 1.94$, $p = .06$), suggesting that on average the guesses of the human judges were better than a chance level. In a second model, we added a binary expertise variable (expert vs. nonexpert) as a fixed effect. There was a marginally significant interaction between expertise and empirically observed effect sizes ($t[475] = 1.88$, $p = .06$), suggesting that experts predictions were marginally closer to empirically observed changes than nonexperts. The variation based on individual judges and individual observation is further depicted in Figure 4. The mixed-effects model analyses conceptually replicated the aggregated-level analysis presented in the preceding paragraph.

In summary, this study showed that predicting effect sizes could be a very difficult task even for experts who have been working on the topic for a period of time. While expertise improved accuracy on individual level, the group-level predictions of experts was only marginally better than the group-level prediction of nonexperts. It is also important to note that the results from this study are somewhat more optimistic than the results observed by Milkman et al. (2021). In our study, we found that human judges are better than chance in predicting intervention efficacy, and, unlike Milkman et al. (2021), we also observed a pattern suggesting that expertise might increase the accuracy of prediction.

Study 3: Comparison of MAB and CED

In the previous two studies, we observed a high degree of heterogeneity of outcomes, and we also found that this heterogeneity is also hard to predict a priori. These were the two conditions that we specified as favoring the application of ML algorithms to select interventions. In this study, we explore how much advantage a MAB algorithm has over CED in terms of time and sample size and to what degree this advantage depends on the amount of expertise.

The hypothetical plot behind the simulations is that researchers have generated 35 interventions and are trying to decide which intervention to launch at scale. On the one hand, they can randomly choose one without testing and launch it. Alternatively, they can start an empirical test using either the CED or the MAB approach, wherein after each participant they revise their belief of which intervention to launch. We also consider cases where the expert uses their expertise to improve the MAB algorithm. The main relationship we are interested in is between the efficacy of the chosen intervention as a function of the number of subjects participating in the empirical test.

Synthetic Data

We used the data collected in Study 1 as an approximation of real-world effect sizes. Each time, we randomly assigned a hypothetical subject to one of the 35 interventions, and then we drew an observed preference shift from a Gaussian distribution with the same mean and variance as those observed in Study 1 for that particular intervention. We ran separate simulations for intervention-level effects and for Intervention × Group-Level Effects.

Role of ϵ

As a MAB algorithm, we used an ϵ-greedy algorithm. We ran a series of simulations where we selected the best intervention after running a given number of subjects. We also varied the level of ϵ between 0.2 and 1 (maximal exploration). We observed that for ϵ between 0.6 and 0.8, we needed about 100 subjects to get very close to the maximally possible shift. For details see Figure 5.

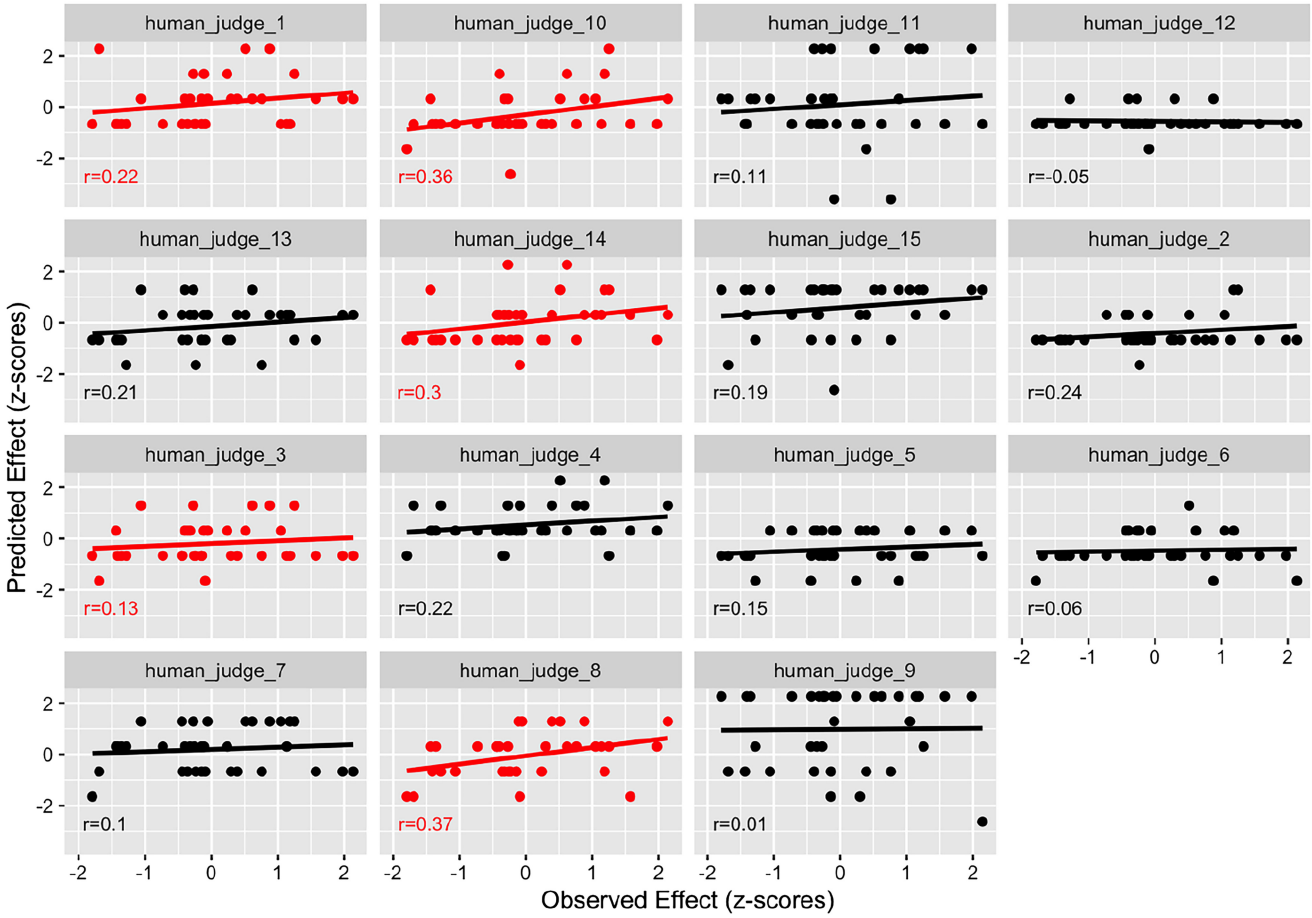
Role of Expertise

Next, we tested to what degree selective expertise can increase the performance of a MAB algorithm. In Study 2, we observed a moderate accuracy of predictions, yet it is possible that in other content domains expertise might lead to more (or less) accurate predictions. It is an open question how much higher expertise accelerates the convergence of a MAB algorithm. For this study, we defined expertise as the correlation between the predicted and the observed effects sizes from a list of interventions.

[8] All five experts had a PhD degree in behavioral science and multiyear experience in running behavioral studies both in academic and industry settings.

Figure 4

Expected Versus Observed Intervention Results for Each Individual Judge



Note. The expected intervention effectiveness is on the *y*-axis, the empirically observed intervention effectiveness is on the *x*-axis. Both scales are *z*-transformed. The data from expert judges are depicted in red and data from nonexpert judges are depicted in black. The lines represent the regression line for each individual judge and the text labels are the Pearson's correlation coefficients.

To check the potential advantage associated with expertise we extended the basic ϵ-greedy algorithm by adding a discounted expertise component. We ranked the interventions based not only on empirically observed effect sizes, but also on the initial expert guesses. For the first simulated subject there is no empirical ranking, so the choice of intervention was entirely based on the expert ranking, and for each additional step s, the role of expert ranking was reduced based on discounting rate r.

$$d = \frac{1}{1 + rs}, \tag{1}$$

and the combined ranking of intervention i is the weighted average of the empirical and expert rankings.

$$\text{combined}_i = \text{expert}_i \times d + \text{empirical}_i \times (1 - d). \tag{2}$$

As an illustration, if the discount rate is $r = 0.1$, at step $s = 10$, then $d = 0.5$. If the expert thought a priori that intervention i should be ranked as $\text{expert}_i = 1$ while after the 10 steps, we observe that intervention i is actually ranked at $\text{empirical}_i = 5$, then the combined ranking for this intervention will be

$$\text{combined}_i = 1 \times 0.5 + 5 \times 0.5 = 3.0. \tag{3}$$

Similarly, if instead of 10, we have 100 steps, then $d = 0.09$

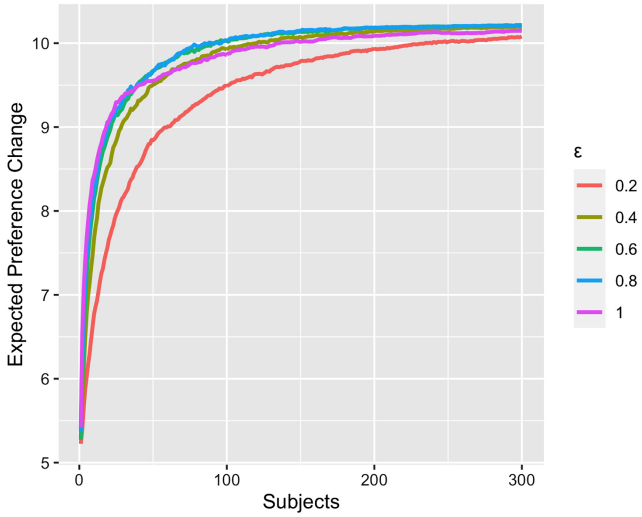
$$\text{combined}_i = 1 \times 0.09 + 5 \times 0.91 = 4.64. \tag{4}$$

The larger the discount rate r is, the faster the empirical ranking takes over the initial expert guess. Additionally, the more data we collect, the lower the role of expertise becomes. While this is a very simple model that might not be very useful in practice without further development, it can still be a helpful tool for understanding the value that expertise might add to a MAB algorithm.

We ran a series of simulations varying the level of expertise and the discount rate. The results are presented in Figure 6. The combination of high expertise and a low discount rate has a substantial advantage at very small sample sizes. However, this advantage quickly disappears at around 50 subjects. Furthermore, maximum effect sizes are achieved after 100 subject regardless the level of expertise.

Figure 5

Expected Intervention Efficacy for Different Sample Sizes and Different Levels of ϵ

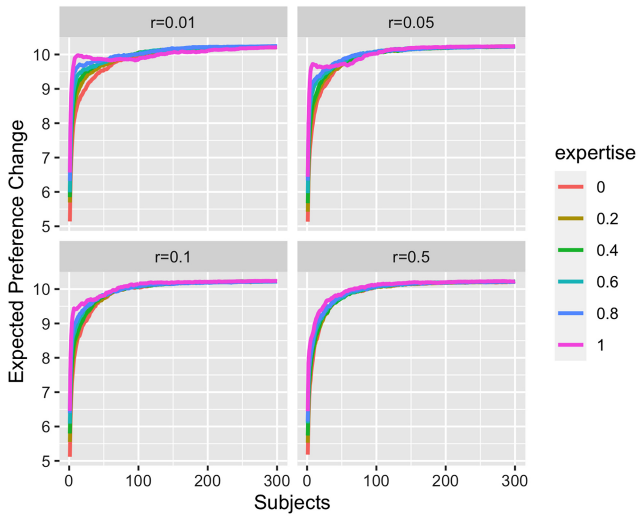


Note. The MAB algorithm aims at selecting the most efficient intervention after observing the results from n subject. The y-axis represents the expected effect size of the currently best intervention based on 500 simulations. The x-axis represents the number of subjects which the MAB algorithm uses to reach that effect size.

In summary, the results from this series of simulations demonstrated that expertise plays a negligible role when the interventions can be tested even with relatively small samples. Medium level expertise (e.g., $r = .4$), as observed in Study 2, is virtually uninformative when empirical testing is available.

Figure 6

A Comparison of Performance at Different Discount Rates (r) *for the Hybrid-MAB Algorithm Which Incorporates Expertise*



Note. Expertise can make a substantial difference (e.g., top left), but only for very small sample sizes. The levels of expertise are measured as the Spearman's correlation between the real and the predicted ranking of interventions. MAB = multiarmed bandit.

Comparison Between CED and MAB Algorithms: Main Effects Only

Next, we compared the expected effect sizes achieved by the best intervention based on different sample sizes using a MAB algorithm and CED. In the classical design, we assumed that we are randomly testing one intervention at a time, using a sample size of 50 people.[9] Similarly to the previous simulations, here we are interested in the expected effect size achieved after running a particular number of subjects. After testing a sequence of interventions, we choose the intervention with highest effect size.

The results are presented in Figure 7. After running a single intervention with 50 people, the expected preference shift is the average of all interventions ($m = 5.33$). In contrast, the expected MAB algorithm's preference shift after 50 people is almost twice as high ($m = 9.67$), and already very close to the expected maximum of $m = 10.29$.

Comparison Between CED and MAB Algorithms: Adding Demographics

The previous simulations were focused exclusively on main effects only. However, in Study 1, we also observed heterogeneity due to demographic factors. An intervention that is the best for one demographic group might not be the best for another demographic group. Therefore, choosing an intervention based on demographic variables can increase the overall impact. In the current simulation, we compare the results of a MAB algorithm for selecting the best intervention given demographic information[10] with a CED approach that tests different interventions for different demographic groups. Again, for the classical design, we assumed a sample size of 50 subjects per intervention per group (e.g., testing all 35 interventions for four demographic groups will take 7,000 subjects). For simplicity, we also assumed that the four demographic groups are equally likely to appear in the population. The results of these simulations are presented in Figure 8. Running 400 subjects, the expected effect size for CED is 5.83, while the expected effect size for MAB is 15.06, which is already very close to the maximum expected shift of 15.93.

Overall, Study 3 leads to two main conclusions. First, there is a large cost advantage of using MAB algorithms over CED. Second, expertise provides little advantage over a MAB algorithm, unless sample sizes are extremely limited and expertise is relatively high.

Discussion

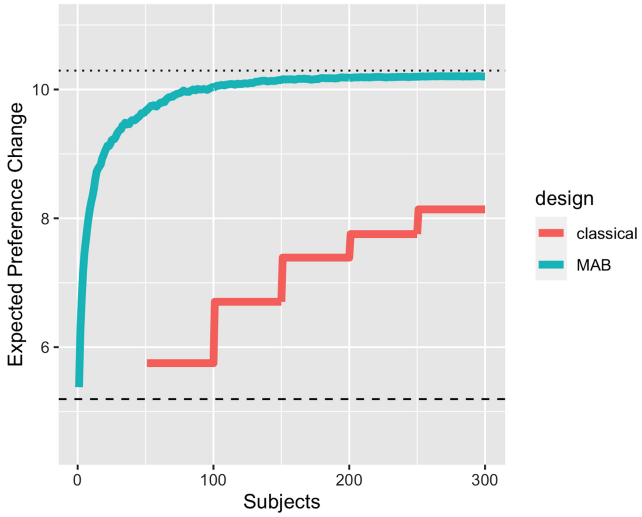
In the 1920s, Vygotsky proclaimed that psychology was in a crisis, and that the root cause of this crisis was the inherent tension between the methodological rigor of academic research and the necessities, constraints, and specific goals of applied research (Dafermos, 2014). Furthermore, Vygotsky claimed that the academic version of psychology has an unjust methodological priority over applied psychology, resulting in developing approaches and theories which do not reflect real world issues. While it is not

[9] A common rule of thumb among behavioral researchers is to use at least 30 subjects per group. A recent work by Brysbaert (2019) suggests at least 50 per group, for example, testing all 35 interventions would require 1,750 subjects.

[10] This problem is often referred to as contextualized bandits.

Figure 7

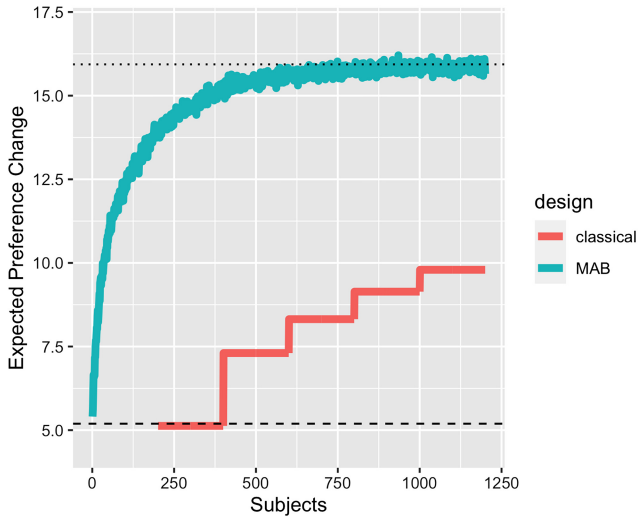
The Expected Best Intervention Effect Size for MAB and CED as a Function of the Sample Size



Note. The dotted line represents the maximal possible effect size and the dashed line represents the expected effect size if we choose an intervention at random and launch it without empirical testing. MAB = multiarmed bandit; CED = classical experimental design.

Figure 8

The Expected Best Intervention Effect Size for MAB and CED as a Function of the Sample Size When Taking Demographic Information Into Account



Note. The dotted line represents the maximal possible effect size and the dashed line represents the expected effect size if we choose an intervention at random and launch it without empirical testing. MAB = multiarmed bandit; CED = classical experimental design.

clear if Vygotsky proposed a viable practical solution to this problem, we agree with the high-level observation that applied research does not need to be a mere application of methods and theories developed in the lab. More specifically, in this article, we argue that the heterogeneity of outcomes needs to be addressed differently in basic versus applied behavioral research. As Linden and Hönekopp (2021) pointed out, in basic research, unaccounted heterogeneity is a major challenge for theoretical advance. However, in applied research, heterogeneity does not need to be necessarily explained. Instead, it can directly be used to maximize desired effects by applying various ML algorithms. To illustrate this point, we explored how information snippets might shift preferences for BEVs. In Study 1, we observed a substantial heterogeneity at the level of main effects and at the level of demographic groups. This finding is closely aligned with the recent work suggesting that heterogeneity of outcomes should be treated as a norm rather than as an exception (Bryan et al., 2021; Linden & Hönekopp, 2021; Milkman et al., 2021). In Study 2, we explored to what degree the observed heterogeneity of main effects is predictable by human judges. While we found that expertise somewhat increases accuracy, even experts who have worked on the topic achieve only modest prediction accuracy. In Study 3, we ran a series of simulations exploring if a basic ML algorithm provides advantages over classical experimental design when heterogeneity is present. The two main findings from this study were that selective expertise matters only when the subject pool is very limited, and that a MAB algorithm provides a huge advantage over CED for small samples. For the same sample sizes, we observed doubling of the expected effect sizes when using a MAB algorithm on level of main effects, and tripling it when taking demographic information into account. In short, the sequence of studies presented here demonstrated that in applied projects, MAB-like approaches can achieve the same effectiveness as CED at a much lower cost (financial, time, and opportunity costs), or can achieve much greater effects for the same cost by including detailed contextual information.

The studies presented here are just a coarse illustration of how ML methods can help behavioral scientists to increase the impact of their applied research. In the real world, a particular applied project might or might not benefit from the approach presented here. Some factors might make those approaches particularly impactful, while other factors will limit the potential advantage we observed here. The method we propose will be particularly useful when:

- allocating participants into suboptimal treatment arms is costly (e.g., backfiring effects),
- creating and implementing a large number of interventions is cost-effective,
- high heterogeneity of outcomes is expected,
- Multiple Group × Intervention categories are possible,
- additional detailed context is available, such as time of day, day of week, social context, or geographical location,
- longitudinal data are available and personalized interventions are possible,
- selective expertise accuracy is medium or low,
- a sequence of interventions are possible,[11]
- interventions can be selected and administered online, and
- the effectiveness of interventions can change over time.

Nevertheless, there are research contexts in which the approach outlined here will be less useful or even not applicable. Two limitations

[11] In this case, MAB algorithms will not be sufficient and full-fledged reinforcement learning approaches will be necessarys.

are worth mentioning. The first one is related to the generalizability of findings. Using ML approaches, particularly those sensitive to context, will be most effective when the research population is the same as the population among which the final intervention launch will be conducted. Although the power of the method we described here is in finding patterns that are not easy to predict from theory or expertise alone, those patterns will most likely be less effective when switching populations or contexts. Second, the methods we described here have statistical properties that are still actively studied. Those properties are less well understood than the statistics used in CED and they are often studied via simulations. For example, statistical power or Type I error rate cannot be inferred for the general case and depends on the particular algorithm being applied, the number of alternatives and the distribution of effect sizes (Villar et al., 2015). Researchers who need accurate estimates a priori would need to apply well-studied algorithms for which such estimates are available, or to run their own simulations.

Conclusion

In the last few years, there has been an increasingly crisp realization that findings from the lab or inferences from psychological theories are difficult to apply to real-world problems. What interventions will work, and what groups or contexts will benefit most from a particular intervention has proved to be difficult to predict a priori. In this article, we propose that a part of the solution for these problems is a greater methodological differentiation between basic and applied behavioral science and we demonstrate that ML can be used to boost selective expertise. Our results suggest that while experts can be better at predicting future outcomes than nonexperts, expertise alone only marginally improves the effectiveness of interventions. Using ML, on the other hand, dramatically improves the effectiveness of the interventions and significantly reduces the required sample sizes, particularly when additional information is present, such as demographic characteristics, personal history or spatio-temporal contexts. While traditional behavioral science methods are still invaluable for advancement of theory, those methods are less effective for selecting real-world interventions (Bryan et al., 2021; DellaVigna & Linos, 2022; Milkman et al., 2021; Yarkoni, 2020). The alternative presented here, on the other hand, is particularly well suited for applied work and while it will not solve the heterogeneity crisis in behavioral science, it will help overcoming some of the real-wold constrains it has posed to researchers.

References

Amadio, B. (2020). *Multi-armed bandits and the stitch fix experimentation platform*. Stitchfix Blog.

Amat, F., Chandrashekar, A., Jebara, T., & Basilico, J. (2018). *Artwork personalization at Netflix*. In Proceedings of the 12th ACM conference on recommender systems (pp. 487–488).

Bouchouicha, R., Deer, L., Eid, A. G., McGee, P., Schoch, D., Stojic, H., Ygosse-Battisti, J., & Vieider, F. M. (2019). Gender effects for loss aversion: Yes, no, maybe? *Journal of Risk and Uncertainty*, *59*(2), 171–184. https://doi.org/10.1007/s11166-019-09315-3

Brückmann, G., Willibald, F., & Blanco, V. (2021). Battery electric vehicle adoption in regions without strong policies. *Transportation Research Part D: Transport and Environment*, *90*, Article 102615. https://doi.org/10.1016/j.trd.2020.102615

Bryan, C. J., Tipton, E., & Yeager, D. S. (2021). Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nature Human Behaviour*, *5*(8), 980–989. https://doi.org/10.1038/s41562-021-01143-3

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, *2*(1), Article 16. https://doi.org/10.5334/joc.72

Burnham, A., Lu, Z., Wang, M., & Elgowainy, A. (2021). Regional emissions analysis of light-duty battery electric vehicles. *Atmosphere*, *12*(11), Article 1482. https://doi.org/10.3390/atmos12111482

Consumer Reports. (2020). *New CR survey finds the majority of consumers are interested in getting an electric vehicle*.

Dafermos, M. (2014). Vygotsky's analysis of the crisis in psychology: Diagnosis, treatment, and relevance. *Theory & Psychology*, *24*(2), 147–165. https://doi.org/10.1177/0959354314523694

DellaVigna, S., & Linos, E. (2022). Rcts to scale: Comprehensive evidence from two nudge units. *Econometrica*, *90*(1), 81–116. https://doi.org/10.3982/ECTA18709

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, Article 621. https://doi.org/10.3389/fpsyg.2015.00621

Gächter, S., Johnson, E. J., & Herrmann, A. (2022). Individual-level loss aversion in riskless and risky choices. *Theory and Decision*, *92*, 599–624. https://doi.org/10.1007/s11238-021-09839-8

Gal, D., & Rucker, D. D. (2018). The loss of loss aversion: Will it loom larger than its gain? *Journal of Consumer Psychology*, *28*(3), 497–516. https://doi.org/10.1002/jcpy.2018.28.issue-3

Goldberg, M. H., Gustafson, A., Rosenthal, S. A., & Leiserowitz, A. (2021). Shifting republican views on climate change through targeted advertising. *Nature Climate Change*, *11*, 573–577. https://doi.org/10.1038/s41558-021-01070-1

Halpern, D. (2015). *Inside the nudge unit: How small changes can make a big difference*. Random House.

Harinck, F., Van Dijk, E., Van Beest, I., & Mersmann, P. (2007). When gains loom larger than losses: Reversed loss aversion for small amounts of money. *Psychological Science*, *18*(12), 1099–1105. https://doi.org/10.1111/j.1467-9280.2007.02031.x

Harinen, T., Filipowicz, A., Hakimi, S., Iliev, R., Klenk, M., & Sumner, E. (2021). *Machine learning reveals how personalized climate communication can both succeed and backfire*. arXiv preprint arXiv:2109.05104

Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods*, *21*(4), 447–457. https://doi.org/10.1037/met0000120

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Horgan, J. (2016). *Psychology's credibility crisis: The bad, the good and the ugly*. Scientific American.

Internal Revenue Service. (2017). *Behavioral science toolkit*.

Jahanbakhsh, K. (2020). *Applying multi-armed bandit algorithms to computational advertising*. arXiv preprint arXiv:2011.10919

Jenn, A. (2020). Emissions benefits of electric vehicles in uber and lyft ride-hailing services. *Nature Energy*, *5*(7), 520–525. https://doi.org/10.1038/s41560-020-0632-7

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 363–391. https://doi.org/10.2307/1914185

King, G. (2011). Ensuring the data-rich future of the social sciences. *Science*, *331*(6018), 719–721. https://doi.org/10.1126/science.1197872

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmertest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Linden, A. H., & Hönekopp, J. (2021). Heterogeneity of research results: A new perspective from which to assess and promote progress in psychological science. *Perspectives on Psychological Science*, *16*(2), 358–376. https://doi.org/10.1177/1745691620964193

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487. https://doi.org/10.1037/a0039400

Milkman, K. L., Gromet, D., Ho, H., Kay, J. S., Lee, T. W., Pandiloski, P., & Park, Y. (2021). Megastudies improve the impact of applied behavioural science. *Nature*, *600*, 478–483. https://doi.org/10.1038/s41586-021-04128-4

Molenaar, P. C., & Campbell, C. G. (2009). The new person-specific paradigm in psychology. *Current Directions in Psychological Science*, *18*(2), 112–117. https://doi.org/10.1111/j.1467-8721.2009.01619.x

Offer-Westort, M., Coppock, A., & Green, D. P. (2021). Adaptive experimental design: Prospects and applications in political science. *American Journal of Political Science*, *65*(4), 826–844. https://doi.org/10.1111/ajps.12597

Pearl, J. (2010). An introduction to causal inference. *The International Journal of Biostatistics*, *6*(2), Article 7. https://doi.org/10.2202/1557-4679.1203

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.

Rau, H. A. (2014). The disposition effect and loss aversion: Do gender differences matter? *Economics Letters*, *123*(1), 33–36. https://doi.org/10.1016/j.econlet.2014.01.020

Schwartz, E. M., Bradlow, E. T., & Fader, P. S. (2017). Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, *36*(4), 500–522. https://doi.org/10.1287/mksc.2016.1023

Shepardson, D., & Klayman, B. (2021). *U.S. government to end gas-powered vehicle purchases by 2035 under Biden order*. Reuters.

Simonson, I., & Kivetz, R. (2018). Bringing (contingent) loss aversion down to earth—A comment on Gal & Rucker's rejection of "losses loom larger than gains". *Journal of Consumer Psychology*, *28*(3), 517–522. https://doi.org/10.1002/jcpy.2018.28.issue-3

Sproul, T., & Michaud, C. P. (2017). Heterogeneity in loss aversion: Evidence from field elicitations. *Agricultural Finance Review*, *77*(1), 196–216. https://doi.org/10.1108/AFR-05-2016-0045

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.

The White House. (2015). *xecutive order no. 13707—Using behavioral science insights to better serve the American people*.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, *25*(3/4), 285–294. https://doi.org/10.2307/2332286

Van Bavel, J. J., Baicker, K., Boggio, P. S., Capraro, V., Cichocka, A., Cikara, M., Crockett, M. J., Crum, A. J., Douglas, K. M., Druckman, J. N., Drury, J., Dube, O., Ellemers, N., Finkel, E. J., Fowler, J. H., Gelfand, M., Han, S., Alexander Haslam, S., Jetten, J., … Willer, R. (2020). Using social and behavioural science to support COVID-19 pandemic response. *Nature Human Behaviour*, *4*(5), 460–471. https://doi.org/10.1038/s41562-020-0884-z

Villar, S. S., Bowden, J., & Wason, J. (2015). Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *30*(2), 199–215. https://doi.org/10.1214/14-STS504

Walton, G. M., & Wilson, T. D. (2018). Wise interventions: Psychological remedies for social and personal problems. *Psychological Review*, *125*(5), 617–655. https://doi.org/10.1037/rev0000115

Wang, M., Rieger, M. O., & Hens, T. (2017). The impact of culture on loss aversion. *Journal of Behavioral Decision Making*, *30*(2), 270–281. https://doi.org/10.1002/bdm.v30.2

Xie, Y., Hwang, S., & Pantelous, A. A. (2018). Loss aversion around the world: Empirical evidence from pension funds. *Journal of Banking & Finance*, *88*, 52–62. https://doi.org/10.1016/j.jbankfin.2017.11.007

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, Article e1. https://doi.org/10.1017/S0140525X20001685

Zajonc, R. B., & Markus, H. (1982). Affective and cognitive factors in preferences. *Journal of Consumer Research*, *9*(2), 123–131. https://doi.org/10.1086/jcr.1982.9.issue-2

Received February 17, 2023
Revision received October 17, 2023
Accepted November 14, 2023 ■