# Promise and Perils of Experimentation: The Mutual Internal Validity Problem

Hause Lin, Kaitlyn M. Werner, and Michael Inzlicht
University of Toronto

Researchers run experiments to test theories, search for and document phenomena, develop theories, or advise policymakers. When testing theories, experiments must be internally valid but do not have to be externally valid. However, when experiments are used to search for and document phenomena, develop theories, or advise policymakers, external validity matters. Conflating these goals and failing to recognize their tensions with validity concerns can lead to problems with theorizing. Experimenters in psychology should be aware of the mutual-internal-validity problem, long recognized by experimental economists. When phenomena elicited by experiments are used to develop theories that, in turn, influence the design of theory-testing experiments, experiments and theories can become wedded to each other and lose touch with reality. They capture and explain phenomena within but not beyond the laboratory. We highlight how triangulation can address validity problems by helping experiments and theories make contact with ideas from other disciplines and the real world.

*Keywords:* theory, experimentation, triangulation, validity, mutual internal validity

In Press at ***Perspectives on Psychological Science***

**Please cite this paper as:** Lin, H., Werner, K. M., & Inzlicht, M. (in-press). Promise and perils of experimentation: The mutual internal validity problem. *Perspectives on Psychological Science*.

"Social science is an example of a science which is not a science. They follow the forms. You gather data, you do so and so and so forth, but they don't get any laws, they haven't found out anything. They haven't got anywhere—yet. Maybe someday they will, but it's not very well developed."

— Richard Feynman in The Pleasure of Finding Things Out (BBC Horizon, 1981, 42:54).

Decades of psychological research has amassed troves of data and produced hundreds of theories (Fiske, 2001). Despite such incredible progress, Richard Feynman and many others' criticisms remain true: Psychological science is messy. Instead of progressing toward establishing a paradigmatic, normal science (Kuhn, 2012; Meehl, 1978; Muthukrishna & Henrich, 2019), psychological theories have become increasingly fragmented (Kruglanski, 2001), and many empirical findings that once provided the basis for many established theories either cannot be replicated or can only be reproduced under narrow conditions (Baker, 2016; Camerer et al., 2018; Open Science Collaboration, 2015; Turner et al., 2018).

Here, we suggest that conflating the goals of experimentation and failing to recognize the tensions between experimentation goals and validity concerns lead to problems with theorizing. Critically, when using experimentation specifically to develop theories, researchers should be aware of the *mutual-internal-validity problem*: Experiments and theories become wedded to each other and lose touch with the real world because experiments are structured according to theories, which are, in turn aimed at describing experimentally-elicited phenomena (Schram, 2005). We then describe how triangulation addresses this problem by bridging multiple theories, methods, and data sources (e.g., Mathison, 1988; Munafò & Smith, 2018). Crucially, we are not suggesting that researchers

**Correspondence** concerning this article should be addressed to Hause Lin, Department of Psychology, University of Toronto, 1265 Military Trail, Toronto, ON, M1C 1A4, Canada. Email: hauselin@gmail.com

abandon laboratory experimentation; instead, we highlight the importance of distinguishing experimentation's goals and recognizing its potential weaknesses, which will better equip researchers to adopt research programs that *enhance* its utility.

## Promises and Goals of Experimentation

Experimentation has been one of the cornerstones of the scientific method since Galileo performed the first recorded laboratory experiment around 400 years ago (Feynman et al., 1963; Pearl & Mackenzie, 2018; Settle, 1961). Increasingly, social scientists like psychologists and economists have also turned to this experimental approach. Further attesting to the importance of this approach comes from the fact that not one, but two Nobel Prizes have been awarded to economists in 2002 and 2019 for their experimental work. Here, we focus exclusively on experiments that aim to maximize control within the laboratory context or other artificial settings (including online experiments) by carefully eliminating extraneous factors that might influence measurements, relationships between variables, and results (Campbell, 1957; Guala, 2003). We do not discuss field experiments that occur in and intervene on natural settings, which often provide less control over extraneous factors (e.g., List & Levitt, 2005).

Experimentation serves various functions and goals. In the last few decades, many scholars, including those in psychology and especially economics, have proposed different taxonomies to capture this diversity (Greenwood, 1982; Roth, 1986; Ribe & Steinle, 2002; Schram, 2005). Broadly, experiments can be designed to (1) test hypotheses or theories (deductive approach), (2) search for and document novel or unexplained phenomena (exploratory and descriptive approach), (3) develop theories (inductive-deductive approach), and, occasionally, (4) advise policymakers (pragmatic approach). Experimental economists have also been especially reflective and receptive to discussions of the need to distinguish these goals and understand how they relate to validity (Guala & Mittone, 2005; List & Levitt, 2005; Loewenstein, 1999; Sugden, 2005). For example, experiments designed to accomplish one goal (e.g., test theories of dopamine function) often have limitations (e.g., limited external validity) that render them less suitable for other goals like searching for and documenting phenomena that exist beyond the laboratory (e.g., when and how dopamine neurons fire).

## Perils of Experimentation

### Tensions Between Internal and External Validity

Threats to internal and external validity have always posed problems to researchers, especially after Donald Campbell recognized and defined these terms over half a century ago (Campbell, 1957; Campbell & Stanley, 1963). Internal validity is the extent to which we can draw confident causal conclusions, whereas external validity is the extent to which we can generalize the conclusions from our experiments to another context. When designing experiments and interpreting findings, the tension between experimentation goals and validity becomes apparent: Experiments provide the most direct way to determine causal effects and test theories because it maximizes control and internal validity by simplifying, isolating, and making tractable even the most complex phenomena (Manzi, 2012; Pearl & Mackenzie, 2018), but these concessions are made at the cost of reducing external validity or the generalizability of the findings.

Unfortunately, unlike economists who have been reflecting on these tensions since the rise of experimental economics in the 1980s (Bardsley, 2005; Hertwig & Ortmann, 2001; List & Levitt, 2005; Loewenstein, 1999; Smith, 1989), many experimenters in psychology have either brushed off these tensions or even defended external invalidity (e.g., Anderson et al., 1999; Mook, 1983), and discussions have been revived only very recently (e.g., generalizability crisis; Yarkoni, 2019).

Decades of debate within experimental economics, however, have led to insights that could be leveraged to improve experimentation and research practices in psychology (Sugden, 2005). Specifically, economists have advised that to properly design and evaluate experiments, researchers should always be aware of the goals of each experiment, which determine the relative importance of internal and external validity (Guala & Mittone, 2005; Schram, 2005). Here, we suggest that conflating the goals of experimentation and failing to recognize the tensions between experimentation goals and validity concerns can also lead to problems with theorizing.

In principle, the ideal experiment is one with high internal and external validity. But in practice, this standard is often unattainable and unnecessary. When the goal is to test theories and deduced hypotheses, internal validity matters much more than external validity because it can be assumed that the hypotheses are expected to be tested under highly specific and all-

other-things-equal conditions (e.g., Guala & Mittone, 2005). For example, to test theories and hypotheses of how context influences subjective value representation and choice (e.g., Gluth et al., 2020; Khaw et al., 2017), one must be able to precisely manipulate variables such as context and subjective value while ensuring other extraneous factors do not bias the causal conclusions. Similarly, theories of anterior cingulate function make predictions that should be tested with experimental paradigms that manipulate and rule out relevant variables (e.g., reward, efficacy, surprise; Frömer et al., 2020; Shenhav et al., 2020; Vassena et al., 2020). In these cases, experimental control is essential.

That is, if the goal is to test theories, experiments should be evaluated by how much they tell us about the underlying theories and not by their resemblance to phenomena in real life (Plott, 1991). For example, when interviewed on the artificiality of experiments and prospect theory's narrow domain, Daniel Kahneman noted that when testing theories, it is acceptable and even normal for experiments and theories to have no relevance to real-life domains (Andersson & Holm, 2002). Therefore, external validity—whether the effects predicted by the theory can be generalized to other domains and real life—is secondary and should be evaluated separately (see Mook, 1983).

However, if the goal is to search for phenomena and document reliable patterns of empirical observations that tell us something general about behavior and mental processes outside the laboratory, researchers will have to consider external validity (Guala & Mittone, 2005; List & Levitt, 2005). For example, to understand how people would respond to moral dilemmas associated with autonomous vehicles (Awad et al., 2018; Bonnefon et al., 2016), it is much more important for the experimental paradigm to reflect the complexities of real-life situations as closely as possible than to ensure it very precisely manipulates specific variables and processes (e.g., moral emotions, mental state representations) and controls for extraneous factors. Without external validity, it would be difficult to determine the value of any empirical observation or regularity, especially if strong theoretical frameworks are also absent.

**Dangers with Combining Experimental Goals**

In practice, researchers often combine the goals of experimentation with insufficient forethought. Researchers design experiments that elicit interesting phenomena, induce underlying causal processes, develop general theories, deduce hypotheses from their theories, and test them by designing experiments that are structured according to the theories. There are two related problems here. The first is the famous problem of induction raised by David Hume centuries ago (for a recent and related perspective in psychology, see Yarkoni, 2019).

The second problem, however, is less obvious, but has been considered a pitfall of experimentation among experimental economists. If experimentally-elicited phenomena are used to develop theories that then shape the design of theory-testing experiments (which again shapes theory development), this repeated, mutual feedback between experiments and theories can lead to a problem with "mutual internal validity" (Guala & Mittone, 2005; Schram, 2005, p. 234). This mutual dependency can lead unwitting experimenters to develop theories that become increasingly capable of explaining phenomena that are "bottled" by the experimental design (Andersson & Holm, 2002; Ross et al., 2010), but no longer describe phenomena outside the laboratory and therefore lose touch with reality (Schram, 2005; Sugden, 2005).

The mutual-internal-validity problem is evident in psychology's most reliable and respected bodies of work. After Tversky and Kahneman experimentally demonstrated how people's decisions are susceptible to inconsequential changes to the decision's context or frame (Tversky & Kahneman, 1981), the observed empirical regularities gave rise to not only prospect theory, but also many popular dual-system theories that suggest that behavior arises from the competition between a fast and emotional system versus a slower and deliberate system (Evans, 2008; Kahneman, 2011).

Subsequent dual-system theorizing and experimentation have taken on such lives of their own (e.g., De Martino et al., 2006; Loewenstein et al., 2015; McClure et al., 2004), that they hint at the mutual-internal-validity problem and raise questions on whether certain theories and phenomena might have been "bottled" by ingenious experimental designs. For example, to test dual-system hypotheses, experimenters have designed laboratory paradigms that force decision-makers to decide under time constraints (e.g., two-second response deadline). Critically, these experiments are structured according to the theory: They assume a priori the existence of dual systems (which is questioned and debated; see Teoh et al., 2020; Melnikoff & Bargh, 2018; Pennycook et al., 2018), and further assume that time-constraint manipulations will purportedly deactivate the slower of the two systems.

Results from this single experimental paradigm have not only inspired many theories (e.g., Rand et al., 2012), but also competing computational models that are capable of explaining highly specific experimental findings (e.g., Chen & Krajbich, 2018; Diederich & Trueblood, 2018; Hutcherson et al., 2015; Teoh et al., 2020). Since these models were specifically designed to describe behavior under time constraint in the laboratory (see also Evans, 2020), it should come as no surprise that they describe phenomena in laboratory contexts very well. However, whether the processes manipulated by this artificial paradigm and described by these models characterize behavior under time constraint outside the laboratory remains unaddressed. That is, the mutual-internal-validity problem has led to increasingly baroque theories whose primary goal is to describe experimentally-elicited rather than real-life phenomena (Schram, 2005). Both the theories and experiments were not only born in the laboratory, but also structured according to each other, and barely touch the outside world they purportedly describe.

Although some might argue that concerns with mutual internal validity and external validity are irrelevant (Mook, 1983), this claim applies only if the goal is solely theory testing, which is extremely rare in practice. Psychology experiments often aim to accomplish multiple goals, and most experimenters would feel embarrassed if they had to admit that their research cannot be extended outside the laboratory walls. Eminent scholars like Daniel Kahneman and Judea Pearl have voiced similar concerns: "good psychology involves a constant interplay between observing the real world and running experiments …, most of my ideas come from the real world, not from the laboratory" (Andersson & Holm, 2002, p. 45) and "scientific progress would grind to a halt were it not for the ability to generalize results …, from test tubes to animals to humans" (Pearl & Mackenzie, 2018, p. 312). Therefore, to prevent "bottled" experiments and theories from mutually reinforcing each other and "losing touch with the natural phenomena" (Tinbergen, 1963, p. 299), researchers should avoid the mutual-internal-validity problem by ensuring their experiments and theories eventually make contact with the real world.

## Theoretical Triangulation Offers Solutions

We suggest that triangulation offers solutions to the mutual-internal-validity problem. The idea of triangulation has been around for many decades (Campbell & Fiske, 1959; Mathison, 1988; Webb et al., 1966). Researchers can triangulate in multiple ways (Denzin, 1978), but they are most familiar with methodological and data triangulation, which address concerns with measurement artifacts arising from using just one method or data source. In fact, psychologists have made tremendous progress toward triangulation by using multi-method approaches (e.g., combine behavior, self-report, neurophysiology, field experiments). Moreover, recent work has also focused on data triangulation across time (e.g., Wang & Inbar, 2020), cultures (e.g., Awad et al., 2018; Ruggeri et al., 2020), investigators, and laboratories (e.g., Moshontz et al., 2018).

Our focus is instead on theoretical triangulation (Denzin, 1978; Mathison, 1988), which can help to remove experiments and theories from their "bottles" to ensure they make contact with ideas from other disciplines and the real world. Thus, it not only directly addresses the mutual-internal-validity problem, but also helps to generate reliable and insightful data that lead to better theoretical integration (Lawlor et al., 2016; Lipton, 2004; Munafò & Smith, 2018). Psychology has already had some success with within-discipline integration, or what we consider as small-interdisciplinary (small-I) triangulation. For example, researchers have bridged social and cognitive psychology with neuroscience, using cognitive neuroscience theories and methods to constrain and inform existing social cognition models as well as inspire new research (e.g., Apps et al., 2016; Lieberman, 2007; Van Lange, 2006).

But the most generative and important research programs often involve big-interdisciplinary (big-I) theoretical triangulation. For example, the theories of evolution and continental drift were informed by ideas from distinct disciplines like paleontology and geology, and many neuroscience and psychology discoveries were possible only by exploiting established principles and theories from physics and engineering. In recent years, many psychologists have begun pursuing ambitious big-I research that has already produced rich data that describe and explain diverse problems like climate change (Pearson et al., 2018), misinformation (Vosoughi et al., 2018), morality, and artificial intelligence (Bonnefon et al., 2016). Although big-I theoretical triangulation is difficult (e.g., Bromham et al., 2016), it is worthwhile because of its potential to improve not just theorizing, but also increase the practical relevance of psychological research (e.g., Bassett & Gazzaniga, 2011; Henrich et al., 2010; Pearson et al., 2014; Pearson et al., 2016; Sloane &

Moss, 2019; Rahwan et al., 2019; van Rooij & Baggio, 2020).

## Beyond Experimentation and Psychology: The Case of Reinforcement Learning

The value of big-I theoretical triangulation is best exemplified by one of psychology's most generative and influential research programs: reinforcement learning. By tracing its history, we show how triangulation had helped this research area flourish, helped it overcome the mutual-internal-validity problem, and improved its theorizing and experimentation.

Since Ivan Pavlov's seminal work on the effects of rewards on organisms' motivation and learning more than a century ago, researchers have been conducting experiments to investigate how the brain responds to and learns from rewards and punishments. Specifically, many theories have been proposed to explain the role of dopamine neurons in reward learning and conditioned behavior (for reviews, see Berridge, 2007). Although a few theories (e.g., dopamine as a pleasure chemical) were overturned by subsequent experimental findings (e.g., Berridge, 2003; Wise, 1996), many competing theories remained and different research groups consistently reported experimental findings that fit with their own theories, which ascribed different roles and functions to dopamine neurons (e.g., wanting versus liking, uncertainty, incentive salience; Berridge, 2007; Fiorillo et al., 2003; Redgrave et al., 1999).

Further experimentation by each research group seemed to only provide evidence that bolstered each group's theory, suggesting problems with mutual internal validity. That is, theories might have been designed to explain the behavior of animals in laboratory experiments (but not in the real world), and experiments and operationalizations of constructs were structured according to these theories. As discussed earlier, these problems cannot be easily addressed by further experimentation (Bardsley, 2005; Schram, 2005), especially if these experiments are similarly structured according to the theories and also lack external validity.

But everything changed in the 90s when researchers fortuitously triangulated. Experimental neuroscientists and psychologists noticed a remarkably close correspondence between dopamine firing patterns and the reward prediction error signal theorized by computer scientists who were trying to develop reinforcement learning algorithms that could learn to perform complex behaviors after exposure to only rewards and punishments (for review, see Niv, 2009). This discovery led to the reward prediction error theory of dopamine (Montague et al., 1996; Schultz et al., 1997). Since then, it has been tested and validated in thousands of experiments with animals and humans (e.g., Eshel et al., 2016; Pessiglione et al., 2006), and prediction errors induced by unexpected outcomes in real life have even been shown to predict real-world behavior (Otto et al., 2016; Villano et al., 2020).

Although the details of this theory are still being debated, it has nevertheless laid the groundwork for the development of subsequent theories and new sub-disciplines. By linking current neuroscience, psychology, and artificial intelligence research, it has inspired many new theories in various disciplines (Dabney et al., 2020; Decker et al., 2016; Dolan & Dayan, 2013; Silver et al., 2017). It has even contributed to new sub-disciplines like neuroeconomics, which then gave rise to distinct frameworks that are nevertheless grounded in the same principles of learning and reward (e.g., Camerer et al., 2005; Glimcher, 2011; Lin & Vartanian, 2018; Padoa-Schioppa, 2011; Polanía et al., 2019; Westbrook & Braver, 2015). This body of psychological work on learning, motivation, and reward has, therefore, highlighted the value of using triangulation to enhance the utility of laboratory experimentation.

## Moving Forward

Although the story of reinforcement learning illustrates the pernicious effects of mutual internal validity, it highlights, more importantly, that research programs can flourish when researchers recognize and directly address the mutual-internal-validity problem. Below, we provide a few suggestions for dealing with it.

The first indication of potential problems with mutual internal validity is the lack of external validity. Research programs and experiments suffering from this problem often produce results that cannot generalize readily (see generalizability crisis; Yarkoni, 2019). Researchers should also be critical of unqualified arguments defending the "external invalidity" of all experiments (e.g., Mook, 1983), because they hold up only when the goal of experimentation is theory testing (e.g., Decker et al., 2020; Vassena et al., 2020). Thus, researchers should evaluate the internal and external validity of an experiment in relation to its goal(s). When designing and preregistering experiments, researchers should also consider explicitly stating the

goal(s) of the experiments and justify the relevance of internal and external validity.

Another potential indicator of mutual internal validity is when theories are developed and tested using a single experimental paradigm. Using only one experimental paradigm risks creating an artificial context in which the experiment and theory co-exist and mutually reinforce, preventing the "bottled" experiment and theory from making contact with or generalizing to the outside world. The example of relying primarily on time-pressure experiments to develop and test dual-system theories highlights how mutual internal validity could creep into even the most rigorous research programs. Other potential examples include implicit cognition theories (Greenwald et al., 1998; Schimmack, 2019), whose development might have over-relied on the implicit association test, and social preference theories that have been based almost exclusively on experiments with dictator games (List & Levitt, 2005).

To address problems with mutual internal validity, experimenters should actively strive to triangulate by using multiple experimental paradigms to test and develop theories. When phenomena and results converge across distinct experimental paradigms, researchers can be more certain that their experiments and theories capture phenomena in the real world, as in the case of research on error- or conflict-related neural potentials (Kumar et al., 2019), which are robustly elicited by different experimental paradigms in distinct contexts and across modalities (Cavanagh et al., 2012; Falkenstein et al., 2000; Lin et al., 2018). But if only one particular experiment can produce theory-consistent phenomena reliably, experimenters whose goals are more than just theory testing should be wary of this "overfitting" of experiment and theory, which hints at the mutual-internal-validity problem.

In addition to using different experiments, researchers should triangulate their methods by using techniques (e.g., cross-validation) from other disciplines that have been designed specifically to overcome overfitting and evaluate whether results and models generalize to unseen or new data (Breiman, 2001; Jordan & Mitchell, 2015; Pearl, 2015). For example, to complement the traditional approach of using t-tests to evaluate whether the means of two experimental groups significantly differ (which overfits the data; see Yarkoni & Westfall, 2017), researchers can consider using a logistic-regression classifier to evaluate how well the trained classifier predicts experimental group assignment on unseen data. This approach provides cross-validation metrics (e.g., prediction accuracy) that tell researchers how well the results generalize and can be presented alongside existing metrics (e.g., probability values, confidence intervals, Bayes factors).

As illustrated by the case of reinforcement learning, although psychology as a discipline awkwardly straddles the hard and soft sciences, this position presents unique opportunities for engaging in big-I theoretical triangulation, which is critical for addressing the mutual-internal-validity problem. To exploit its position, the discipline should strive to provide more well-rounded training that focuses more on breadth of knowledge and researchers should be encouraged (by collaborators, reviewers, editors) to explicitly engage with work from diverse disciplines ranging from applied engineering research to theoretical ideas from biology and sociology. For example, in the introduction and discussion sections of manuscripts, researchers can make efforts to consider how their constructs and theories fit with those in related disciplines or use theoretical approaches from other disciplines to inform the methods used in psychological research (e.g., Grahek et al., 2020; Mosleh et al., 2020; Mosleh et al., forthcoming; Rahwan et al., 2019). Many research programs are already engaging in big-I triangulation, but the discipline can do even more to exploit its position.

## Conclusion

Experimentation can help to test theories, search for and document phenomena, develop theories, and, occasionally, advise policymakers. Tensions exist between these goals and validity issues, and failing to recognize these tensions can lead to problems with theorizing. Internally valid experiments are ideal for theory testing, whereas externally valid experiments matter more for the other goals. Crucially, experimenters in psychology should be aware of the mutual-internal-validity problem recognized by experimental economists. When researchers use experiments to search for phenomena as well as develop and test theories, theories can become increasingly capable of explaining phenomena within the lab but lose touch with reality (Schram, 2005).

Many problems with experimentation can be addressed by triangulation, which provide multiple ways to improve scientific inference and generalizability (e.g., Pearl & Mackenzie, 2018). Triangulation can help to identify not only convergent and valid findings but also inconsistent and contradictory ones that need to be made sensible and

coherent (Mathison, 1988). Crucially, it is more than just using multiple methods and data sources—it is an approach that provides more and better evidence from which researchers can construct reliable and psychological theories that cohere with established bodies of theories (e.g., Krugman, 2014; Meehl, 1967; von Hayek, 1974), as well as evolving cultural and historical contexts (e.g., Henrich et al., 2010; Muthukrishna & Henrich, 2019; Schulz et al., 2018).

Like many other disciplines, psychology is facing a crisis. But it also presents opportunities to improve how research is done (Brock, 2019; Editorial, 2020). As the discipline reflects, it can again, look elsewhere for inspiration. Meteorology was considered an abysmal science decades ago while it struggled with the messiness of weather forecasting (Gleick, 2008), but data, method, and theoretical triangulation have completely revolutionized the discipline.

## References


Anderson, C. A., Lindsay, J. J., & Bushman, B. J. (1999). Research in the psychological laboratory: Truth or triviality. *Current Directions in Psychological Science*, *8*(1), 3-9. https://doi.org/10.1111/1467-8721.00002

Andersson, F. N., & Holm, H. (2002). *Experimental economics: Financial markets, auctions, and decision making: Interviews and contributions from the 20th Arne Ryde Symposium*. Springer Science & Business Media.

Apps, M. A., Rushworth, M. F., & Chang, S. W. (2016). The anterior cingulate gyrus and social cognition: Tracking the motivation of others. *Neuron*, *90*(4), 692-707. https://doi.org/10.1016/j.neuron.2016.04.018

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J. F., & Rahwan, I. (2018). The Moral Machine experiment. *Nature*, *563*(7729), 59-64. https://doi.org/10.1038/s41586-018-0637-6

Baker, M. (2016). *1,500 scientists lift the lid on reproducibility*. Retrieved 2020-02-21 from https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

Bardsley, N. (2005). Experimental economics and the artificiality of alteration. *Journal of Economic Methodology*, *12*(2), 239-251.

Bassett, D. S., & Gazzaniga, M. S. (2011). Understanding complexity in the human brain. *Trends in Cognitive Sciences*, *15*(5), 200-209.

BBC Horizon. (1981). *Richard Reynman: The pleasure of finding things out [Video file]*. Retrieved 2019-12-22 from Retrieved from https://www.dailymotion.com/video/x6ptg1x

Berridge, K. C. (2003). Pleasures of the brain. *Brain and Cognition*, *52*(1), 106-128.

Berridge, K. C. (2007). The debate over dopamine's role in reward: the case for incentive salience. *Psychopharmacology (Berl)*, *191*(3), 391-431. https://doi.org/10.1007/s00213-006-0578-x

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573-1576. https://doi.org/10.1126/science.aaf2654

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199-231.

Brock, J. (2019). *"It's not a replication crisis. It's an innovation opportunity"*. Retrieved 2020-02-21 from https://www.natureindex.com/news-blog/not-a-replication-crisis-innovation-opportunity

Bromham, L., Dinnage, R., & Hua, X. (2016). Interdisciplinary research has consistently lower funding success. *Nature*, *534*(7609), 684-687.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., . . . Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637-644.

Camerer, C., Loewenstein, G., & Prelec, D. (2005). Neuroeconomics: How neuroscience can inform economics. *Journal of Economic Literature*, *43*(1), 9-64. https://doi.org/10.1257/0022051053737843

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*(4), 297. https://psycnet.apa.org/record/1959-03494-001

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Houghton Mifflin Company.

Cavanagh, J. F., Zambrano-Vazquez, L., & Allen, J. J. (2012). Theta lingua franca: A common mid-frontal substrate for action monitoring processes. *Psychophysiology*, *49*(2), 220-238.

Chen, F., & Krajbich, I. (2018). Biased sequential sampling underlies the effects of time pressure and delay in social decision making. *Nature Communications*, *9*(3557), 1-10. https://doi.org/10.1038/s41467-018-05994-9

Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C. K., Hassabis, D., Munos, R., & Botvinick, M. (2020). A distributional code for value in dopamine-based reinforcement learning. *Nature*, *577*(7792), 671-675. https://doi.org/10.1038/s41586-019-1924-6

De Martino, B., Kumaran, D., Seymour, B., & Dolan, R. J. (2006). Frames, biases, and rational decision-making in the human brain. *Science*, *313*(5787), 684-687. https://doi.org/10.1126/science.1127205

Decker, A., Finn, A., & Duncan, K. (2020). Errors lead to transient impairments in memory formation. *Cognition*, *204*, 104338.

Decker, J. H., Otto, A. R., Daw, N. D., & Hartley, C. A. (2016). From creatures of habit to goal-directed learners: Tracking the developmental emergence of model-based reinforcement learning. *Psychological Science*, *27*(6), 848-858. https://doi.org/10.1177/0956797616639301

Denzin, N. K. (1978). *The research act: A theoretical introduction to sociological methods and practice*. Sage.

Diederich, A., & Trueblood, J. S. (2018). A dynamic dual process model of risky decision making. *Psychological Review*, *125*(2), 270-292.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312-325.

Editorial. (2020). Irreproducibility is not a sign of failure, but an inspiration for fresh ideas. *Nature*, *578*(7794), 191-192. https://doi.org/10.1038/d41586-020-00380-2

Eshel, N., Tian, J., Bukwich, M., & Uchida, N. (2016). Dopamine neurons share common response function for reward prediction error. *Nature Neuroscience*, *19*(3), 479-486. https://doi.org/10.1038/nn.4239

Evans, J. S. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review of Psychology*, *59*, 255-278.

Evans, N. J. (2020). Think fast! The implications of emphasizing urgency in decision-making. https://doi.org/10.31234/osf.io/pfrb4

Falkenstein, M., Hoormann, J., Christ, S., & Hohnsbein, J. (2000). ERP components on reaction errors and their functional significance: A tutorial. *Biological Psychology*, *51*(2), 87-107.

Feynman, R. P., Leighton, R. B., & Sands, M. (1963). *The Feynman lectures on physics*. Addison-Wesley Publishing Company.

Fiorillo, C. D., Tobler, P. N., & Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, *299*(5614), 1898-1902. https://doi.org/10.1126/science.1077349

Fiske, S. T. (2001). Social psychology, theories of. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences* (pp. 14413-14421). Pergamon.

Frömer, R., Lin, H., Dean Wolf, C. K., Inzlicht, M., & Shenhav, A. (2020). When effort matters: Expectations of reward and efficacy guide cognitive control allocation. *bioRxiv*. https://doi.org/10.1101/2020.05.14.095935

Gleick, J. (2008). *Chaos: Making a new science*. Penguin Books.

Glimcher, P. W. (2011). *Foundations of neuroeconomic analysis*. Oxford University Press.

Gluth, S., Kern, N., Kortmann, M., & Vitali, C. L. (2020). Value-based attention but not divisive normalization influences decisions with multiple alternatives. *Nature Human Behaviour*, *4*(6), 634-645.

Grahek, I., Musslick, S., & Shenhav, A. (2020). A computational perspective on the roles of affect in cognitive control. *International Journal of Psychophysiology*, *151*, 25-34.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464. https://osf.io/zye9u/download

Greenwood, J. D. (1982). On the relation between laboratory experiments and social behaviour: Causal explanation and generalization. *Journal for the Theory of Social Behaviour*, *12*(3), 225-250.

Guala, F. (2003). Experimental localism and external validity. *Philosophy of Science*, *70*(5), 1195-1205. https://doi.org/10.1086/377400

Guala, F., & Mittone, L. (2005). Experiments in economics: External validity and the robustness of phenomena. *Journal of Economic Methodology*, *12*(4), 495-515. https://doi.org/10.1080/13501780500342906

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world. *Behavioral and Brain Sciences*, *33*(2-3), 61-83.

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists. *Behavioral and Brain Sciences*, *24*(3), 383-403. https://doi.org/10.1037/e683322011-032

Hutcherson, C. A., Bushong, B., & Rangel, A. (2015). A neurocomputational model of altruistic choice and its implications. *Neuron*, *87*(2), 451-462.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255-260. https://doi.org/10.1126/science.aaa8415

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Khaw, M. W., Glimcher, P. W., & Louie, K. (2017). Normalized value coding explains dynamic adaptation in the human valuation process. *Proceedings of the National Academy of Sciences*, *114*(48), 12696-12701. https://doi.org/10.1073/pnas.1715293114

Kruglanski, A. W. (2001). That "vision thing": The state of theory in social and personality psychology at the edge of the new millennium. *Journal of Personality and Social Psychology*, *80*(6), 871-875.

Krugman, P. (2014). *The dismal science*. Retrieved 2019-08-19 from https://www.nytimes.com/2014/09/28/books/review/seven-bad-ideas-by-jeff-madrick.html?_r=0

Kuhn, T. S. (2012). *The structure of scientific revolutions*. University of Chicago Press.

Kumar, A., Gao, L., Pirogova, E., & Fang, Q. (2019). A review of error-related potential-based brain–computer interfaces for motor impaired people. *IEEE Access*, *7*, 142451-142466.

Lawlor, D. A., Tilling, K., & Davey Smith, G. (2016). Triangulation in aetiological epidemiology. *International Journal of Epidemiology*, *45*(6), 1866-1886.

Lieberman, M. D. (2007). Social cognitive neuroscience: A review of core processes. *Annual Review of Psychology*, *58*, 259-289.

Lin, H., Saunders, B., Hutcherson, C. A., & Inzlicht, M. (2018). Midfrontal theta and pupil dilation parametrically

track subjective conflict (but also surprise) during intertemporal choice. *NeuroImage*, *172*, 838-852. https://doi.org/10.1016/j.neuroimage.2017.10.055

Lin, H., & Vartanian, O. (2018). A neuroeconomic framework for creative cognition. *Perspectives on Psychological Science*, *13*(6), 655-677. https://doi.org/10.1177/1745691618794945

Lipton, P. (2004). *Inference to the best explanation*. Psychology Press.

List, J. A., & Levitt, S. D. (2005). What do laboratory experiments tell us about the real world. *NBER working paper*, 14-20.

Loewenstein, G. (1999). Experimental economics from the vantage-point of behavioural economics. *The Economic Journal*, *109*(453), 25-34. https://doi.org/10.1111/1468-0297.00400

Loewenstein, G., O'Donoghue, T., & Bhatia, S. (2015). Modeling the interplay between affect and deliberation. *Decision*, *2*(2), 55-81.

Manzi, J. (2012). *Uncontrolled: The surprising payoff of trial-and-error for business, politics, and society*. Basic Books.

Mathison, S. (1988). Why triangulate? *Educational Researcher*, *17*(2), 13-17.

McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science*, *306*(5695), 503-507. https://doi.org/10.1126/science.1094492

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, *34*(2), 103-115.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806-834.

Melnikoff, D. E., & Bargh, J. A. (2018). The mythical number two. *Trends in Cognitive Sciences*, *22*(4), 280-293. https://doi.org/10.1016/j.tics.2018.02.001

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, *16*(5), 1936-1947.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, *38*(4), 379-387.

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., Grahe, J. E., McCarthy, R. J., Musser, E. D., & Antfolk, J. (2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, *1*(4), 501-515. https://doi.org/10.1177/2515245918797607

Mosleh, M., Kyker, K., Cohen, J. D., & Rand, D. G. (2020). Globalization and the rise and fall of cognitive control. *Nature Communications*, *11*(1).

Mosleh, M., Pennycook, G., Arechar, A. A., & Rand, D. G. (forthcoming). Twitter data reveal digital fingerprints of cognitive reflection. *Nature Communications*. https://doi.org/10.31234/osf.io/qaswn

Munafò, M. R., & Smith, G. D. (2018). *Robust research needs many lines of evidence*. Retrieved 2020-02-17 from https://www.nature.com/articles/d41586-018-01023-3

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*, 221-229. https://doi.org/10.1038/s41562-018-0522-1

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139-154. https://doi.org/10.1016/j.jmp.2008.12.005

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Otto, A. R., Fleming, S. M., & Glimcher, P. W. (2016). Unexpected but incidental positive outcomes predict real-world gambling. *Psychological Science*, *27*(3), 299-311. https://doi.org/10.1177/0956797615618366

Padoa-Schioppa, C. (2011). Neurobiology of economic choice: A good-based model. *Annual Review of Psychology*, *34*, 333-359.

Pearl, J. (2015). Generalizing experimental findings. *Journal of Causal Inference*, *3*(2). https://doi.org/10.1515/jci-2015-0025

Pearl, J., & Mackenzie, D. (2018). *The book of why*. Penguin UK.

Pearson, A. R., Schuldt, J. P., & Romero-Canyas, R. (2016). Social climate science: A new vista for psychological science. *Perspectives on Psychological Science*, *11*(5), 632-650. https://doi.org/10.1177/1745691616639726

Pearson, A. R., Schuldt, J. P., Romero-Canyas, R., Ballew, M. T., & Larson-Konar, D. (2018). Diverse segments of the US public underestimate the environmental concerns of minority and low-income Americans. *Proceedings of the National Academy of Sciences*, *115*(49), 12429-12434. https://doi.org/10.1073/pnas.1804698115

Pearson, J. M., Watson, K. K., & Platt, M. L. (2014). Decision making: The neuroethological turn. *Neuron*, *82*(5), 950-965.

Pennycook, G., Neys, W. D., Evans, J. S. B. T., Stanovich, K. E., & Thompson, V. A. (2018). The mythical dual-process typology. *Trends in Cognitive Sciences*, *22*(8), 667-668. https://doi.org/10.1016/j.tics.2018.04.008

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*(7106), 1042-1045.

Plott, C. R. (1991). Will economics become an experimental science. *Southern Economic Journal*, *57*(4), 901-919. https://doi.org/10.2307/1060322

Polanía, R., Woodford, M., & Ruff, C. C. (2019). Efficient coding of subjective value. *Nature Neuroscience*, *22*(1), 134-142. https://doi.org/10.1038/s41593-018-0292-0

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D.,

McElreath, R., Mislove, A., Parkes, D. C., Pentland, A., . . . Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477-486. https://doi.org/10.1038/s41586-019-1138-y

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, *489*(7416), 427-430.

Redgrave, P., Prescott, T. J., & Gurney, K. (1999). Is the short-latency dopamine response too short to signal reward error. *Trends in Neurosciences*, *22*(4), 146-151. https://doi.org/10.1016/S0166-2236(98)01373-3

Ribe, N., & Steinle, F. (2002). Exploratory experimentation: Goethe, Land, and color theory. *Physics Today*, *55*(7), 43-49.

Ross, L., Lepper, M., & Ward, A. (2010). History of social psychology: Insights, challenges, and contributions to theory and application.

Roth, A. E. (1986). Laboratory experimentation in economics. *Economics and Philosophy*, *2*, 245-273. https://doi.org/10.1017/s1478061500002656

Ruggeri, K., Alí, S., Berge, M. L., Bertoldo, G., Bjørndal, L. D., Cortijos-Bernabeu, A., Davison, C., Demić, E., Esteban-Serna, C., Friedemann, M., Gibson, S. P., Jarke, H., Karakasheva, R., Khorrami, P. R., Kveder, J., Andersen, T. L., Lofthus, I. S., McGill, L., Nieto, A. E., . . . Folke, T. (2020). Replicating patterns of prospect theory for decision under risk. *Nature Human Behaviour*, *4*(6), 622-633. https://doi.org/10.1038/s41562-020-0886-x

Schimmack, U. (2019). The implicit association test: A method in search of a construct. *Perspectives on Psychological Science*, *28*(4), 1-19.

Schram, A. (2005). Artificiality: The tension between internal and external validity in economic experiments. *Journal of Economic Methodology*, *12*(2), 225-237.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593-1599. https://doi.org/10.1126/science.275.5306.1593

Schulz, J. F., Barahmi-Rad, D., Beauchamp, J., & Henrich, J. (2018). The Origins of WEIRD Psychology. https://doi.org/10.31234/osf.io/d6qhu

Settle, T. B. (1961). An experiment in the history of science. *Science*, *133*(3445), 19-23.

Shenhav, A., Musslick, S., Botvinick, M. M., & Cohen, J. D. (2020). Misdirected vigor: Differentiating the control of value from the value of control. https://doi.org/10.31234/osf.io/5bhwe

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., & Hassabis, D. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), 354-359.

Sloane, M., & Moss, E. (2019). AI's social sciences deficit. *Nature Machine Intelligence*, *1*(8), 330-331.

Smith, V. L. (1989). Theory, experiment and economics. *Journal of Economic Perspectives*, *3*(1), 151-169. h

Sugden, R. (2005). Experiments as exhibits and experiments as tests. *Journal of Economic Methodology*, *12*(2), 291-302. https://doi.org/10.1080/13501780500086248

Teoh, Y. Y., Yao, Z., Cunningham, W. A., & Hutcherson, C. A. (2020). Attentional priorities drive effects of time pressure on altruistic choice. *Nature Communications*, *11*(1), 3534.

Tinbergen, N. (1963). On aims and methods of ethology. *Zeitschrift für tierpsychologie*, *20*(4), 410-433.

Turner, B. O., Paul, E. J., Miller, M. B., & Barbey, A. K. (2018). Small sample sizes reduce the replicability of task-based fMRI studies. *Communications Biology*, *1*, 62. https://doi.org/10.1038/s42003-018-0073-z

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science*, *211*(4481), 453-458.

Van Lange, P. A. M. (2006). *Bridging social psychology: Benefits of transdisciplinary approaches*. Psychology Press. https://doi.org/10.4324/9781410616982

van Rooij, I., & Baggio, G. (2020). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science.

Vassena, E., Deraeve, J., & Alexander, W. H. (2020). Surprise, value and control in anterior cingulate cortex during speeded decision-making. *Nature Human Behaviour*, *4*, 412-422.

Villano, W. J., Otto, A. R., Ezie, C. E. C., Gillis, R., & Heller, A. S. (2020). Temporal dynamics of real-world emotion are more strongly linked to prediction error than outcome. *Journal of Experimental Psychology: General*.

von Hayek, F. (1974). *Lecture to the memory of Alfred Nobel: The pretence of knowledge*. Retrieved 2019-12-22 from https://www.nobelprize.org/prizes/economic-sciences/1974/hayek/lecture/

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146-1151.

Wang, S. Y., & Inbar, Y. (2020). Moral language use by U.S. political elites. *Psychological Science*.

Webb, E. J., Campbell, D. T., & Schwartz, R. D. (1966). *Unobtrusive measures*. Chicago: Rand McNally.

Westbrook, A., & Braver, T. S. (2015). Cognitive effort: A neuroeconomic approach. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(2), 395-415.

Wise, R. A. (1996). Addictive drugs and brain stimulation reward. *Annual Review of Neuroscience*, *19*(1), 319-340.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100-1122. https://doi.org/10.1177/1745691617693393

Yarkoni, T. (2019). The generalizability crisis. https://doi.org/10.31234/osf.io/jqw35

This pre-print was designed using the following template: Wiernik, B. M. (2019, October 11). Preprint templates. https://doi.org/10.17605/OSF.IO/HSV6A