The Psychology of Construal in the Design of Field Experiments*

Elizabeth Levy Paluck & Eldar Shafir

Department of Psychology &
Woodrow Wilson School of Public and International Affairs
Princeton University

March 20, 2016

*Revised version, prepared following the NBER Conference on Economics of Field Experiments, organized by Esther Duflo & Abhijit Banerjee. Thank you to Robin Gomila, Sachin Banker, Peter Aronow, and Ruth Ditlmann for helpful comments. Address correspondence to epaluck@princeton.edu and eshafir@princeton.edu.

1 Introduction

Why might you be interested in this chapter? A fair assumption is that you are reading because you care about good experimental design. To create strong experimental designs that test people's responses to an intervention, researchers typically consider the classically recognized motivations presumed to drive human behavior. It does not take extensive psychological training to recognize that several types of motivations could affect an individual's engagement with and honesty during your experimental paradigm. Such motivations include strategic self-presentation, suspicion, lack of trust, level of education or mastery, and simple utilitarian motives such as least effort and optimization. For example, minimizing the extent to which your findings are attributable to high levels of suspicion among participants, or to their decision to do the least amount possible, is important for increasing the generalizability and reliability of your results.

Psychologists agree that these motivations are important to consider when designing experiments, but they rank other behavioral drivers higher. Some drivers of individual behavior often ignored by other experimental researchers, which psychologists consider critical, include: consistency, identity, emotional states like pride, depression, and hunger, social norms, and the perception of notions like justice and fairness. Moreover, psychologists are keenly aware of features of the immediate situation that promote or diminish these behavioral drivers. The question for any experimenter is: how do we figure out which behavioral drivers matter in any one particular experiment, and how they matter?

In this chapter we focus on the notion of construal, an under-appreciated concept that psychologists employ to understand behavior and to design experiments that can better approximate and

help isolate the causal dynamics that lead to the behavior of interest. Construal is defined as an individual's subjective interpretation of a stimulus, whether the stimulus is a choice set, a situation, another person or group or people, a survey, or an experimental intervention (Ross and Nisbett, 1991). Thus, for example, an individual's construal of various items in a survey will depend on the subjective meaning that she attaches to the survey as a whole. She may interpret a survey about risk preferences as a "survey about whether I would be a good investor," even if that is not how the survey is introduced by the investigator, nor what the investigator is really trying to understand.

In the last two decades, psychological insights have been integrated into the study of economic perception and behavior, creating a sub-discipline called behavioral economics or applied behavioral science (Kahneman, 2013). One result is that many economists interested in behavior have a greater appreciation for the seemingly mundane immediate situational features that can promote or diminish behavioral drivers and thus the behaviors themselves. Most behaviorally-informed scientists from a range of disciplines can now tell you that the "situation matters." For example, removing small demands on a person's time, such as signing a form or altering defaults, can dramatically increase take-up rates, like signing up for 401K plans or becoming an organ donor (Thaler and Sunstein, 2008).

The interdisciplinary behavioral science literature has generated a great deal of advice on how to design an intervention, given that the situation matters – advice to the effect that individuals are sensitive to the timing, physical location, milieu, and framing of an intervention (Datta and Mullainathan, 2014; Shafir, 2013). And while psychology has been merged with economics to create a more "behaviorally-informed science", psychologists have pointed out that this field would never be fully "behavioralized" (Ross and Nisbett, 1991). In other words, when studying behavior we

cannot ignore individuals' subjective thoughts about the behavior, much as the early psychologists tried to do when studying Stimulus-Behavior response patterns by training rats and pigeons to respond to lights and sounds (Skinner, 1960; Seligman, 1970). This is because the interventions and research designs we create are never interpreted directly – as they are – or as the experimenter might have intended. Instead, our interventions and research tools are "construed" in ways that must be understood to know what participants are actually responding to.

There is, in this sense, a presumption in standard economic thinking that is really quite radical from the point of view of a psychologist. Economic theorizing posits that people choose between options in the world: Job A versus Job B, or, if one is looking to buy a car, Cars A, B, or C, where the decision maker takes into consideration the information at their disposal. As it turns out, however, what people decide about, or choose between, are not options are they are in the world but, rather, as they are mentally represented. When a person is presented with a choice between options A and B, she chooses not between A and B as they are in the world, but rather as they are represented by the 3-pound machine behind her eyes and between her ears. And that representation is not a complete and neutral summary, but rather a specific and constructed rendering – a construal.

Building on previous work that discusses how to design interventions based on an understanding of situational pressures and individuals' construal of those pressures (Datta and Mullainathan, 2014; Ross and Nisbett, 1991; Shafir, 2013), this chapter points out ways in which participants' construal of your experiment – everything from the behavior in question to the setting, the intervention, the deployment of the intervention, and the measurement tools – should affect the way you design and deploy your experiment.

Acknowledging subjective interpretation of the experiment is not the same as claiming to have no knowledge of participants' construction of reality. Psychologists can provide many ways in which construal processes might be systematic and predictable. Nonetheless, one deep message is that experimenters need to be modest about and to explicitly test assumptions concerning how participants view experimental interventions. Being aware of and taking steps to understand participants' construal in advance can help you to design and deploy the kind of field experiment that will shed light on the causal processes leading to the behavior in which you are interested.

You as the investigator, furthermore, are not excluded from the forces of construal. Toward the end of this chapter, we will also explore how your own construal of your experiment and of the data can affect the way you interpret your results, conduct replications, and recommend elements of your intervention for scaling up or for institutional policies. We begin by providing an overview of construal: its definitions, functions, and some illustrative examples.

Principle of construal

At the turn of the 20th century and particularly during the two world wars, psychologists were moving away from a Freudian focus on personal histories and individual differences driving behavior and behavioral disorders. Kurt Lewin, a German psychologist and an èmigrè who eventually directed the Center for Group Dynamics at MIT, developed a situationally-driven alternative to Freud's claim that conflicting forces within the individual (the id, the ego, and superego), only available through the introspection of the individual and her therapist, could explain behavior and individual decision making. To facilitate the scientific study of behavior, Lewin proposed, we

should look for conflicting forces in the environment surrounding the individual, such as laws, family pressures, social norms, peers, and even the physical environment, and analyze how those forces push an individual and her self-proclaimed beliefs and desires into particular behavior choices.

Lewin called these conflicting forces in the environment, pushing and pulling an individual's behavioral choices, a tension system. The tensions he mapped were between individual motivations and environmental forces. Through a series of field experiments, Lewin showed how leaders, workplace hierarchies, peers' public behavior, and the physical proximity of particular resources could promote or inhibit a person's personal desires and beliefs, and change behavior in predictable ways (Lewin and Gold, 1999). His early theorizing formed the foundation of modern social and cognitive psychology, and today guides the assumptions that psychologists make as they design and evaluate of behavior change and decision making experiments.

As Lewin was exploring the importance of situational pressures on behavior, some psychologists took this view to the extreme, including radical behaviorists like B.F. Skinner who felt that all behavior was no more than a response to objective environmental forces learned over time. This view, while at first popular, proved profoundly insufficient. Particularly glaring was the absence of a principle of human thought that Lewin proposed as a critical part of the tension system analysis: construal. Environmental forces were not directly and objectively perceived by the individuals inhabiting a tension system, Lewin reasoned. Perception is a subjective process, which can happen in a considered, deliberate fashion or in a fast and less conscious manner. Construal, the act of interpreting and attaching subjective meaning to forces like one's peers, leaders, workspace, group identities, choices, and the like, is also inherently variable – a stimulus may be interpreted by the same person one way at a certain time or in a certain situation, and differently in the next situation.

Similarly, two people experiencing the same stimulus can construe it in different ways.

Some classic examples of how construal can affect judgments and behavior include the following:

- Judgments of a stimulus depend on how you construe the judgment relative to similar stimuli you have adapted to in your environment: A rule is perceived as strict when you are used to lax rules, and as lax when you are coming from a stricter rule environment. This is intuitive, and easily demonstrated through a comparable physical experience, that of judging water temperature with your hand, just after you have plunged your hand first in a cold or a hot bucket of water. Judgment will be relative and not reflective of an absolute physical (or social) property (Weber et al., 1996).

- Framing affects construal: Framing a monetary amount as a loss or a gain changes its construal, and the risk attitude it elicits (Tversky and Kahneman, 1981). More generally, any frame depicting a stimulus (an idea, choice, or behavior) as consistent with or as a departure from a perceived reference point shifts an individual's reception of the stimulus (Kahneman and Miller, 1986).

- Self-appraisal is made through social comparisons: Judgments about the self, including accomplishments, motivations, the strength of particular identities, and ideologies, are often made relative to other individuals present in the situation or other individuals mentioned in the question (Markus and Kunda, 1986; Morse and Gergen, 1970).

- Taxes and subsidies provoke unintended reactions, depending on individuals' construal of the behaviors they target: Individuals may construe economic incentives as psychological

taxes (i.e., demotivating) when the incentives subsidize behavior that is self-motivated – the small economic reward replaces what before was a substantial psychological boost. Likewise, economic taxes may be interpreted as psychological subsidies (motivating) when they are imposed on behavior that individuals have mixed feelings about or are trying to stop – the original feeling of guilt is alleviated through the fine (Miller and Prentice, 2013).

- Peer pressure is effective not just because of conformity but because peers redefine the behavior in question: Individuals do not just adopt peer behavior, but also their peers' construal of the behavior or the situation. For example, when individuals observe peers ranking "politician" very positively vs. very negatively as a profession, the individual's own ranking of the term politician changes, not out of mimicry but because the individual has a different kind of politician in mind as a result of their peers' ranking (Asch, 1940).

- Global judgments color more specific ones and earlier information changes the meaning of later information: For example, global traits such as warmth can change the construal of a more specific trait like intelligence: the latter is interpreted as wisdom when a person is globally judged to be warm but as cunning when the person is thought of as cold. Along similar lines, learning about a teacher's argument with a student is interpreted differently if it is first vs. later revealed that the teacher was voted teacher of the year by his students (Ross and Nisbett, 1991).

- The source of a message colors the meaning of the message: Asch showed in a classic study (Asch, 1948) that the quote "a little rebellion now and then is a good thing" was interpreted significantly differently by students when it was attributed to Thomas Jefferson vs. Lenin. Moreover, the words in a message can color the message. When asked about the wisdom of

potential US intervention in a foreign crisis, Americans report significantly different levels of endorsement depending on whether the situation – otherwise described in identical terms – uses a few words (blitzkrieg invasion, troop transports) associated with WWII, or else words (quickstrike invasion, Chinook helicopters) reminiscent of the Vietnam war (Gilovich, 1981).

- Ideology changes which facts are noticed, believed, and remembered: Partisanship determines which facts individuals attend to, believe, remember, and understand when consuming news or other kinds of fact-based reports (Vallone et al., 1985).

- Construal affects how individuals assess the relative importance of various causal factors. While lay people (and researchers) reasonably search out various types of "data" to understand the causes of behavior in the world, including observations of distinctiveness (how specific is the behavior to this instance or individual), consistency (over time, is this behavior observed in this situation or for this individual), and consensus (how many other people behave this way or in this situation), people are often biased toward dispositional explanations of behavior that focus on a person's character, over situational explanations of that behavior that rely on the pressures of the environment (Kelley, 1973; Ross and Nisbett, 1991).

In the words of the cognitive psychologist Jerome Bruner, individuals who construe stimuli differently according to current levels of adaptation, frames, social comparisons, and present desires are "going beyond the information given" (Bruner, 1957). Psychologists see this subjective interpretation as a normal feature of human cognition, which can happen deliberately and consciously as well as spontaneously and unconsciously. That construal can be an automatic and unconscious process troubles our ability as investigators to ask directly about how an individual's interpretation might depend on their current circumstance. Indeed, individuals do not usually have insight into

Figure 1: Three cars of identical size

the ways in which problem presentation, peers, and other Lewinian environmental pressures affect their own construal.

The elements that influence the construal of social circumstances can be subjective and subtle. Let us provide, therefore, one final example from visual perception, where the elements are more objective and clear. Consider the picture below. If you the experimenter were to present this picture to a person in a study, you would be showing them cars of identical size (on paper), but the person would be construing cars of very different sizes. If you asked the participant how much magic marker would be needed to cover each of the three cars in the picture, you would get different estimates, despite the fact that they are identical in size. Here, of course, cues of depth and perspective generate – imperceptibly, effortlessly, universally – a construal of different car sizes. (In fact, in this example, you the experimenter, if you didn't know better, might rely on those cues too.) The factors underlying the construal of social contexts are more varied and less obvious, but they similarly generate subjective representations that depart from what was "objectively presented."

Fortunately, psychologists have identified some "systematic factors [that contribute] to vari-

ability and instability of meaning" in people's construal processes (Ross and Nisbett, 1991, 69). Ross and Nisbett (1991) review the classic literature on various "tools of construal," which include knowledge structures like scripts, schemas, models, and various heuristics that help individuals to quickly and with minimal effort make sense of other people, situations, choices, and assorted stimuli.

Schemas, for example, are mental constructs representing knowledge about a group of related topics. Once a schema is activated, subsequent stimuli are interpreted along the lines suggested by that schema, with consequences for memory, decisionmaking, judgment, and behavior. A schema for "farm," for example, will influence an individual's attention when considering a farm environment; she would spend more time paying attention to aspects of the farm that do not fit with her farm schema, like the appearance of an octopus. In this case, her schema will predict what she expects to see, what piece of information about the farm she spends the most time considering, as well as what she remembers about the farm (Loftus and Mackworth, 1978). Scripts, such as a script for how to behave at an academic conference, contain even more specific knowledge structures about the order in which certain events should unfold and how an individual is expected to behave during each event, such as a discussion section, a coffee break, or an evening dinner with colleagues (Schank and Abelson, 2013).

Scripts, schemas, and heuristics (Gilovich et al., 2002) may be investigated as local tools of construal that exist within certain populations (such as among people in farms or at academic conferences), or as tools of construal that apply to most people (such as the status quo bias against change, which manifests itself in many different populations (Eidelman and Crandall, 2009; Kahneman et al., 1991)). These various tools of construal improve individuals' ability to interpret novel

situations, and even if they sometimes guide behavior and judgments in directions that deviate from the predictions of rational actor models, they help to make resulting behaviors and judgments more fluid and more predictable.

Laypeople and social science researchers often fail to appreciate the role of construal in guiding people's responses; instead, they tend to attribute choices that deviate from some rational prediction or norm to individuals' dispensational characteristics like intelligence, personality, or ideology. The literature on construal encourages the view that behavior is not necessarily a product of a person's character, but rather a window into how the person construes their choices or environment. "Where standard intuition would hold the primary cause of a problem to be human frailty, or the particular weakness of a group of individuals, the social psychologist would often look to situational barriers and to ways to overcome them" (Ross and Nisbett, 1991).

For experimenters (and for policy makers, the consumers of much of this research) this insight should be of great importance. Behavior in experiments, and its interpretation, is determined not simply by the objective building blocks of the experiment, but by what participants know, want, attend to, perceive, understand, remember, and the like. Thus, experiments that are otherwise well designed, including well-intentioned interventions, can fail because of the way they are construed by the participants, or by the investigators themselves. The difference between success and failure can sometimes boil down to a relatively benign and supposedly immaterial change in presentation and subsequent construal, rather than a complex and costly rearrangement of experimental logic or procedure.

In the following pages we show how psychologists understand construal as important to the

design of an effective experiment. We offer a number of suggestions for how you as an investigator can attempt to understand your participants' construal of the stimuli in your experiment, or how you might reach what we term *shared construal* with your participants. The goal is to design and deploy a stimulus (intervention) in a field setting that participants will construe the way you intend them to.

By shared construal we do not mean that investigators and participants understand a behavioral problem or a choice set in the same way. Naturally, the experimenter (say, a professor studying children's candy preferences) may view the options very differently than do the subjects. The experimenter will also know things that the participant does not, and might arrange things in ways that escape the participant's attention. What we mean by shared construal is that the investigator inhabits participants' perspective as best they can, as they are designing the experiment. When Mischel (Mischel et al., 1972) designed his famous marshmallow experiments, he needed to know that kids found those strange, almost non-food-like treats, which he may have detested, irresistible.

Psychologists think of designing experiments as a way of creating different counterfactual worlds for their participants to inhabit and respond to. As the saying goes, "I can explain it for you, but I can't understand it for you." The point, then, is to design a world in which a participant understands the world in the way the experimenter intended, and without her having to explain too much of it. How to do this is no easy feat, and there is no foolproof recipe to follow. In the following sections we offer suggestions for understanding participants' construal, as well as your own, as you conceptualize your intervention and experiment (piloting phase), as you design and deploy your intervention and measurement (design phase), and as you interpret your results and plan follow up experiments or scale ups (interpretation phase).

2 Pilot: seek shared construal of behavior and the situation between investigators and participants

Piloting often means testing out an experimental paradigm before the actual trial. But piloting can also be time set aside to understand a participant population's construal of the behavior in question and of the situations involved in your experimental paradigm prior to designing the full experiment. In this sense, piloting is an investigation and discovery stage about construal. It requires a high level of modesty, curiosity, and openness on the part of the experimenter, to better understand what is driving people, how they see the problem in question, before crafting an intervention to test their behavior.

Before designing the intervention or the experimental paradigm (i.e., the content or the set-up and deployment of the manipulation and measurement), it is important to first understand the underlying drivers of the behavior in question in the particular setting of interest. What are the "restraining forces" that cause the behavior not to be enacted, or the "compelling forces" that drive the behavior at particular times or among particular people?

Redelmeier et al. (1995) were interested in why homeless adults in a southeast region of Toronto, Canada, repeatedly visited the emergency room for care of non-life-threatening ailments, up to 60 times per year, even when they were not given everything they needed. A common construal among medical professionals and researchers was that the behavior was driven by the homeless adults' neediness and the appeal of a warm place of shelter, and that if hospitals provided more care, this would only increase demand. The authors used survey data to understand the construal among

homeless adults who attended the ER: many reported being treated rudely by hospital staff, and nearly half reported that their needs were not met at the time of their visit. Crucially, 42% reported that they returned to the ER because of an unmet medical need.

Based on this alternate construal, Redelmeier et al. (1995) hypothesized that increased care might address homeless adults' perceived satisfaction with their care experience, and lessen the number of return visits. This informed their experimental design: a compassionate care condition run by volunteers who provided randomly-assigned homeless adults with extra (though non-clinical) attention during their visit through friendly conversations and other kinds of rapport-building, and a baseline condition in which the other half of the selected sample were treated as per emergency room policy. In this case, the compassionate treatment, which directly addressed participants' construal of the situation, led to a 30% drop in repeated visits to the emergency room.

It is notable that the experiment excluded homeless adults who might be unresponsive to changes in treatment, including those who were acutely psychotic, unable to speak English, or were intoxicated or extremely ill. These choices, along with the insights regarding participants' vs. the medical professionals' understanding of the triggers of repeated visits to the hospital, were won through familiarity with the context of the experiment, a willingness to admit uncertainty in the standard interpretation of the observed behaviors, along with some data collection regarding the participants' own construal.

Investigating participants' construal of the behaviors of interest ahead of the experiment may shift the intervention design, helping you to re-conceptualize what is at issue for your experiment. The emergency room experiment is one in which individuals with a "big picture" view of the sit-

uation, the hospital administration and medical professionals, had the wrong construal. Piloting helped to uncover a different insight into the behavior, a point that is also made by the literature on intervention design (Datta and Mullainathan, 2014)). The lesson of the emergency room study is not only that the intervention achieved the right construal, but, more importantly, that highly experienced hospital professionals had the wrong construal all along. This is remarkable also because of what was needed for the revision of construals: ask the clients what they thought and felt.

Piloting to understand local construals can help craft the most effective control or comparison conditions. It can help create the most salient contrast that draws out the causal factor believed to be responsible for the behavior under investigation. In the Redelmeier et al Redelmeier et al. (1995) experiment, experimenters did not offer financial incentives, for example, but rather focused on the way clinical treatment was delivered – with compassion.

Piloting can also help you to understand more about participants' construal of the environments where you plan to conduct your experimental manipulations or measurements. The choice of an intervention site ought not to be guided by logistical convenience alone (though this is often critical to the successful deployment of a field experiment). Psychological research on context effects suggests that the site of the experiment can often drive some aspect of the experimental results, a point often less appreciated compared to other concerns. Obviously, you spend lots of time designing the form of your experimental intervention – say, a community meeting, versus a phone-call, or a letter. Once designed, will you convene your community meeting in a church, in an old school, or in a restaurant? Will you send your thoughtfully crafted letter, or make the phone call, to a person's home, or to their workplace? At the beginning or the end of the month?

By this point, it will not surprise you that psychologists believe these choices matter deeply for how your participants will construe your intervention and the issues addressed by it. In the famous Milgram obedience study, participants were ordered by an experimenter to apply (ultimately fake) electric shocks to another study participant when he failed at a memory task. In the version of the study run at Yale University, 65% of participants were fully obedient to the experimenter's commands in delivering the maximum level of shock; 48% of participants were fully obedient when the study was run at a nondescript office building in the nearby city of Bridgeport without a visible university affiliation (and nearly nobody obeyed when the instructions were conveyed by phone) (Milgram, 1974).

Consider also a study of context and behavior by Berger et al. (2008), who examined voting outcomes when voters were assigned to vote in churches vs. schools. First, using observational data, they estimated that voters were approximately .5 percentage points more likely to vote in favor of increasing education spending (by raising the state sales tax from 5.0% to 5.6%) when they had been assigned to vote in a school vs. a church. Second, using an experiment in which participants were initially shown images of either schools or office buildings before stating their policy preference, the authors suggested that the school context primed participants to think positively about education and to vote in its favor. This effect held even though none of the participants believed that that exposure to school images boosted their support for the increased sales tax to support education, "suggesting environmental stimuli can influence voting choice outside of awareness" (p. 8847).

This study highlights an important tension. An experiment's piloting stage is the right time to worry about things like the unintended effects of the context, or the under-appreciated perspectives participants bring with them as one aims to achieve shared construal. At the same time, it is im-

portant to keep in mind that participants are unlikely often to be the most useful informants. After all, if participants had good insight into what drives their behaviors, we could simply ask them – no need to run expensive studies. As it turns out, construal processes are mostly out of contact with conscious awareness. By running carefully controlled studies we can find regularities in people's construal, of which the participants themselves are largely unaware.

3 Design: ensure the intervention design, measurement, and deployment achieve shared construal between investigators and participants

3.1 Intervention design and deployment

Do participants in your field experiment understand the content of your intervention in the same way that you do as the investigator? In a now classic study, Gneezy and Rustichini (2000) introduced fines for picking up children late from daycare in a random subset of a sample of daycare centers in Israel. A fine is normally understood as a deterrent to action, and we might predict that parents in the treatment daycares would be motivated to show up on time, given the increased economic costs to their delay.

Instead, it appears that parents perceived the fines to be what some psychologists have termed an "economic tax but a psychological subsidy" (Miller and Prentice, 2013). Parents in daycares

where fines were implemented were significantly *more* likely to pick up their children late, an effect that persisted even after the fine was removed 17 weeks later. Gneezy and Rustichini (2000) and others have reasoned that the fine reshaped the parents' understanding of their environment. In particular, the contract between parents and daycare providers changed regarding pickups. The fine clarified the contract – picking up your child late "costs" this amount of money. So parents willing to pay the price came late. Another way of stating these results is that parents initially construed on-time pickups as a moral imperative; being late meant you were violating it. The fine was thus construed as a psychological subsidy, a release from this moral guilt. Parents released from this moral obligation now felt they only had to pay, and no longer experienced guilt about a late pickup.

What about community members and other bystanders to your experimental intervention? One negative externality of a field experiment might be that other (non-targeted) people in your participants' social networks may construe the intervention in unintended ways, and influence your participants. Ross and Nisbett (1991) describe the surprising results of the Cambridge Somerville study, in which at-risk boys were randomly assigned to receive or not to receive a bucket of treatments for an extended period of time during early adolescence, including after school and summer programming, tutoring, home visits, and more. In the forty-year follow up to the experiment, investigators found that treatment participants had no better outcomes than control participants, and in some aspects including adult arrests and mortality, treatment participants looked somewhat worse.

Ross and Nisbett (1991) reason that one potential explanation for this lack of observed response to treatment rests in the community's construal of and response to the intervention. For example, community members like coaches and ministers who might have naturally reached out to the at-risk boys may have perceived that the treated boys no longer needed the help of the community, and

withdrew crucial support. Another possibility is that community members construed the treated boys as much worse "troublemakers" due to all the outside attention that they received, and treated them as such. These are post hoc proposals, but plausible ones that remind us of the importance of understanding the community's construal – even when the community is not directly implicated in the experimental manipulation, particularly because they might affect the actual findings.

Anticipating different construals, and achieving shared construal of your intervention design and the way it is deployed[1] in the participant population and the surrounding community is no small task. The examples we used point to the necessity of running a small-scale version of the intervention to invite reactions and reported construals of the intervention that are not merely hypothetical in nature. In cases like the emergency room, interview those experiencing the treatment. Or, as in the case of the daycare experiment, interview parents to see how they understand their current "contract" with the daycare – what drives them to come late, and how do they think the daycare feels about late pickups. Only when parents' construal of the late fee was understood could researchers explore an intervention predicated on a shared construal, which would yield the desired reduction as opposed to increase in late arrivals.

Finally, although it arrives after the implementation of the intervention, all experiments should involve some form of a manipulation check, which assesses whether and what the participant understood and noticed about the intervention. Manipulation checks are used all the time in psychological experiments, for descriptively understanding how participants perceived the intervention, but they are relatively rare outside of psychology. Manipulation checks can be much more than a simple determination of treatment delivery, for generating the estimated LATE given randomized intention

[1]See also recent work by Haushofer and Shapiro (2013) on participant construal of the fairness of the process of random assignment.

to treat. They can give a picture of the participants' construal of the intervention, through questions like "what did the letter tell you?" or "who sent that letter, and why do you think they sent it?" after participants are sent letters about, say, an opportunity for financial literacy training. When we run "deception-free" studies, we might want to inquire whether participants actually fully believe it. And when we tell them a treatment assignment is "random," might participants suspect it is actually rigged? We have encountered cases where participants were convinced an attractive experimenter was "part of the study." And others where parts of the study – claims such as "we are genuinely interested in your beliefs and preferences" – were dismissed. More intrusive manipulation checks via surveys or interviews can happen for a small subsample of the target population, or during piloting.

3.2 Measurement of outcomes and processes

How do participants construe your measurement tools? Do they understand your survey questions the way they were meant to be understood? Do community members assisting with an archival data collection (e.g., photos of a neighborhood over time) perceive the data collection to be appropriate, and do they share the investigator's belief that the records of interest represent accurate traces of the behaviors under study?

Although survey measures are considered second-best to unobtrusively measured behavioral outcomes, they are often desirable additional pieces of information or the only source of outcome measurement in institutionally weak or disorganized settings without good records of behavior. Fortunately, an enormous literature in psychology on psychometrics, heuristics, and biases provides a framework for thinking about when participants' construal of survey questions may differ from

that of the investigator's.

When participants read or listen to a series of questions, they often engage in active acts of interpretation, or misinterpretation. They do not merely listen, or read, and then respond. Rather, they try, often quite innocently, to interpret what it is the investigator is looking for, what is meant by each question in light of the previous question. Questions are not handled in isolation, and a general attempt is made to make global sense of the questionnaire, assessing its general purpose and its broad themes. For a striking example, one of the most widely-used questionnaires in psychology is the Rosenberg Self-Esteem Scale, which features a series of survey items aimed at assessing an individual's self-esteem – none of which include the term *self-esteem*. Participants rate their agreement with items such as "On the whole, I am satisfied with myself," and "All in all, I am inclined to feel that I am a failure" (reverse scored).

Robins et al. (2001) intuited that participants taking this scale would quickly construe the purpose of the scale to be the measurement self-esteem, and that a response to a direct question about self-esteem would be equally valid. They constructed an alternative questionnaire consisting of one item: "I have high self-esteem." Ratings of this single item correlated to the same degree as did the multi-item self-esteem questionnaire with a broad number of criterion measurements, including other self-evaluations and biases, mental and physical health, and peer ratings of the participant. The single-item survey also cut down on the number of complaints from participants about answering the same question multiple times, and reduced the number of skipped questions or random responses and other problems with the multi-item survey protocol.

To be fair, in many cases a more complex topic necessitates multiple items; our point here

is not that surveys must be short, but that participants are not passive recipients of each survey item. Their interpretations, of course, may overlap to various degrees with the investigator's own understanding. Many psychologists use the technique of "cognitive interviewing" (Willis, 2004) to test participants' understanding of a questionnaire before broader deployment. This technique involves asking the participant to react aloud to each question, talking through their reaction to the question, also in light of responses to preceding questions, and why they are providing the responses they provide.

Participants can also construe certain questions in meaningfully different ways, simply as a result of what comes to mind as a function, for example, of the ordering of questions. Schwarz and Xu (2011) inquired about drivers' enjoyment commuting to work in luxury as opposed to economy cars. In one study, they asked University of Michigan faculty and staff which car they drove (brand, model, and year) and subsequently, how they "usually" feel while commuting. Consistent with common intuition, drivers reported more positive emotions when they drove more luxury cars. Thus, estimated mean scores for drivers' positive affect while commuting was significantly higher while driving cars corresponding to the Bluebook values of a BMW than of a Honda Accord.

A reversed order of questioning, however, paints a different picture. In this ordering, university faculty and staff were first asked to report how they felt during their most recent episode of driving to work, and only then after they had reported their feelings, were they asked what car they drove. In this condition, the quality of the car driven, as indexed by (the natural log of) its Bluebook value, was thoroughly unrelated to the drivers' affective experience.

These and similar findings make a simple but important point: What is momentarily on people's

mind can influence their construal. The car matters to reported judgments of enjoyment when it is on the driver's mind, but not otherwise. When asked to report how they usually feel while commuting, drivers who are led to think about their car, arrive at answers that correlate with its value. But when the car goes unmentioned, its value figures not at all.

In other cases, participants respond to a slightly different question, or perform a slightly different computation, than that requested by the investigator, particularly when the concepts involved are only superficially understood by the participants. Item substitution is a phenomenon that was observed in the classic Linda-type problems Tversky and Kahneman (1973)[2] gave participants a description of a fictitious graduate student shown along with a list of nine fields of graduate specialization. Here is a description:

> Tom W. is of high intelligence, although lacking in true creativity. He has a need for order and clarity, and for neat and tidy systems in which every detail finds its appropriate place. His writing is rather dull and mechanical, occasionally enlivened by somewhat corny puns and by flashes of imagination of the sci-fi type. He has a strong drive for competence. He seems to have little feel and little sympathy for other people and does not enjoy interacting with others. Self-centered, he nonetheless has a deep moral sense.

One group of participants was given a representativeness (or similarity) question; others were given a probability question. Participants in a representativeness group ranked the nine fields of specialization by the degree to which Tom W. "resembles a typical graduate student" in each of those fields. Participants in the probability group ranked the nine fields according to the likelihood of Tom W. specializing in each. Figure 1 below plots the mean judgments of the two groups. The correlation between reported representativeness and probability is nearly perfect (.97), showing

[2] see also Kahneman and Frederick (2002), for further discussion.

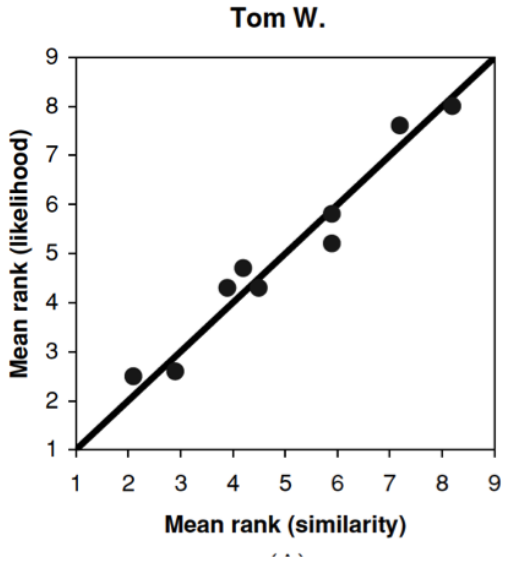

Figure 2: Tversky and Kahneman, 1973

near-perfect attribute-substitution. Representativeness judgments – which are natural and automatic – are more accessible than probability judgments, which are not intuitive and can be rather difficult. (And there is no third attribute that could easily explain both judgments.) When asked about probability, a concept at once subtle yet familiar enough not to require further clarification, people substitute similarity judgments for their response. This, of course, can lead to actual error, where things that are more similar, but less likely, are rated higher in likelihood. (The study also showed that participants' probability judgments correlated highly negatively with their own estimated base rates of the graduate fields of specialization.)

Probability is an example of a concept that feels familiar and straightforward, yet generates responses based on other considerations, like similarity or fear, which have little to do with actual probability. Along similar lines, one needs to worry about what it is exactly that respondents are responding to, what precisely are they computing, when asked about concepts such as anger, or depression, or wellbeing. Many investigators who work with less educated populations, for example, use pictures to help with participant construal of the questionnaire – quite literally, pictures to illustrate the point. Naturally, participants use subjective interpretation with pictures as well as with words, so it is important to pilot how well those pictures are able to communicate the intended question or response options. One of us used a pictorial scale of depression for a field experiment conducted in Rwanda. The scale had been previously used in published work in the same country and more broadly in the Great Lakes region of Africa. It asked participants to answer the question "how have you been feeling in the past few weeks" by pointing to one of a series of pictures featuring a person carrying a stone. From picture to picture, the stone increased in size: on one end of the scale, the person held a small stone in his hand, and at the other end of the scale, the person was bent in half as they held up the weight of an enormous boulder on their shoulders.

Because the scale had been used successfully in previous studies in the area, we brought the scale directly into the field without a pilot. At one site, a participant was asked how they were currently feeling, and was shown the pictorial scale. The participant waited, and then left the interview to confer with others nearby. When he returned, he informed the interviewer that he was willing to carry some of the smaller stones for him, but not some of the larger ones. The misunderstanding of the scale ran even deeper. The scale caused active discussions in this community, and we were informed that during the recent civil conflict a military group asked a group of young men from the community to help carry supplies for them, and the young men were never seen again. A scale to

measure depressive reactions to trauma was construed by the community as related to one of their original sources of trauma. We took care to clarify our intentions and to repair the situation, threw out the scale from our study, and resolved never again to use a scale without a pilot.

3.3 Investigator presence

How do participants construe who you are, as an investigator, and what your presence in their community means for them and for their participation in the experiment? Some ethical discussions encourage investigators to stay away from certain data collections or intervention deployments because participants' respect for or fear of scientists may lead them to construe participation or responses to certain types of questions as mandatory (Orne, 1962; Rosnow and Rosenthal, 1997).

Paluck (2009) reports that varying levels of government scrutiny and physical security in the post-conflict countries where she has deployed field experiments has led to different self-presentation strategies for interviewers and other representatives of the experiment. For example, in Rwanda, where security was excellent and government scrutiny was extremely high, research staff identified themselves strongly with the university supporting the investigator and the study. However, just across the border in the Democratic Republic of Congo (DRC) where security and government surveillance were low, staff wore t-shirts featuring the local NGO that was collaborating with the university. In Rwanda, participants would have construed the emphasis on the NGO to mean that their responses were subject to government surveillance, as were most NGOs in the country during the experiment. However, in DRC, participants needed reassurance of legitimacy from a known local source, the NGO, due to the lack of security, and did not construe the NGO as an actor that

would share their answers with the government.

Many other examples are possible, but our bottom line is that the perceived source of the experiment will affect participants' construal of their choice to participate or not, the confidentiality of their responses, and the overall meaning of the experiment, among other things. Of course, a "social desirability bias" – the tendency to answer questions in a manner that will be viewed favorably by others, in this case by the experimenters – is a serious risk as well. (Social desirability bias can be somewhat alleviated via the use of self-administered computer surveys, and an attempt at highly neutral question wording.) We may even use the analogy of your own construal of the source of information in this chapter: as an economist reading this chapter, might you find certain aspects of it more or less authoritative if you knew they were coming from two economists, sociologists, psychologists?

4 Interpret: how do investigators construe what matters in the data?

Thus far, our focus has been on participants' construal. But investigators use the same tools of construal as participants: we construe what participants do and what they tell us in ways that may or may not match up with their actual actions or meanings. We resort here to the cartoonist Gary Larson for illustrations (see Figures 3 and 4). In the first, we can think of participants as the birds, with their own view of our role and purpose, and of us, strolling in the park, interacting in ways we fail to appreciate. In the second, we may find ourselves in the role of Larson' dog Ginger, hearing

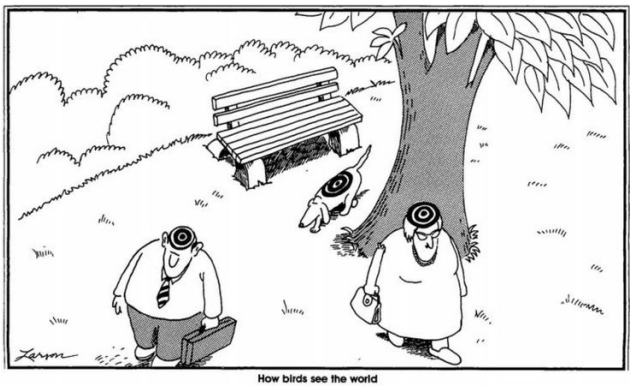

Figure 3: Investigators stroll in the park. Copyright Gary Larson.

Figure 4: Investigators listening to participants. Copyright Gary Larson.

only the information we deem relevant, as participants tell us things we fail to decipher.

Construing participants' self reports is not the only way that construal processes operate for investigators, and can shape the way they understand experimental outcomes. Construal can also affect the ways in which we conduct data analysis, and the factors we interpret to be important for a replication or for scaling up an intervention.

Recently, social scientists have laid out a rationale and evidence for the advantages of preregistration of analyses prior to the deployment of a field experiment or to the commencement of analysis (Casey et al., 2011; Olken, 2015; Committee, 2015). Just as (Vallone et al., 1985) pointed out that partisanship can affect what individuals see in a factual news article, so too can researchers selectively pick analyses that support their preferred hypothesis in a large dataset (Casey

et al., 2011). As Olken puts it, "Even researchers who have the noblest of intentions may end up succumbing to the same sorts of biases when trying to figure out how, ex-post, to make sense of a complex set of results" (p. 1).

Psychologists understand this practice as a result of the ordinary and sometimes inevitable process of construal: what you understand to be the most important test at the design stage can change as you observe the process of data collection, as you analyze your data, and as you form, or perhaps slightly revise, a working hypothesis about the study results. While there are non-negligible costs to pre-registering all of your analyses in advance (Olken, 2015) there are also clear advantages. In addition to publicly committing to a priori predictions, preregistration can help investigators think more carefully about their hypotheses as they design and modify the experimental protocol. A similar practice that can help the post hoc downweighting of experimental hypotheses is pre-registering a field experiment. This practice helps to prevent the selective reporting of entire trials that do not yield the results expected by investigators (using, for example, `http://www.socialscienceregistry.org/` or the newly-instantiated Open Science Framework).

Construal can also influence which factors investigators take to be the generalizable lesson of the overall experiment: i.e., it can shape what is seen as the causal driver of the results. At first blush, this may seem counterintuitive. Randomization of an independent variable, after all, allows for the estimation of a causal relationship. But how do investigators interpret what exactly was the important feature of the independent variable, in order to replicate or to scale up their study?

Consider the field experiment conducted in South Africa, in which Bertrand et al. (2010) ma-

nipulated information a bank provided about loans in letters to their clients. manipulated the information a loan provider included in letters sent to their clients offering financial loans. Some of the information was central to what clients should want to know about the loan terms, including size of loan, duration, and interest rates. Other "information" was peripheral, such as various examples of possible loans one could take, or a picture of a man's vs. a woman's face, embedded in the letter's graphic design. As predicted, the researchers found that some peripheral features in the letter had substantial impact on loan take-up. For example, for male customers, having a picture of a woman on the letter significantly increased demand for the loan, about as much as dropping the interest rate 4.5 percentage points, a reduction of about 25% of the loan interest rate.

How to interpret this experiment? What precisely does it show? Does it show that pictures of women especially increase take-up of loans? Should we always expect pictures of women's faces to increase the take-up of financial products? Would pictures of women be equally effective in Belgium, or would other kinds of pictures prove more effective there? How investigators construe the role played by the woman's picture as the causal driver of loan take-up in their experiment determines how they might try to replicate the experiment in other contexts, or how they might want to institutionalize or scale-up their results for the specific bank or other banks with which they work in South Africa or elsewhere. Replicating experiments in slightly different contexts, such as different banks or governments or other firms, introduces the possibility that participants will construe the intervention differently. This is particularly relevant when the intervention might be perceived as originating from a very different source, or might be associated with slightly different constructs. "Women," after all, like many other possible peripheral cues may play different roles, carry different symbolic connotations, and have different association with financial markets in different places.

Replicating experiments

At its most general level, the South Africa loan experiment teaches us that simple, seemingly peripheral tweaks to advertisements of financial products can make a big difference. Beyond that, it may be unclear how to construe the specifics of the manipulation, for example, regarding the role played by the woman's photo. It is the investigator's challenge to attempt to distill what was most important – and likely to remain stable – about the original significant result for the attempt at replication or scale-up.

Our advice is to think about the conversion of specific manipulations in an experiment like you would about the conversion of currency. Shekels will work well for you in Israel, both in Tel Aviv and in Jerusalem, but it would be a mistake to try to "replicate" that in Japan. Similarly, a manipulation that has worked well in one place ought to work well in another that shares the necessary common features, but it may well fail when transported to a context that differs in some important ways. Because of construal, this advice may be a bit less obvious, or easy to apply, than might first seem. Both field and lab experimental replications are often based on replicating the surface structure – the Shekels – without replicating the deeper structure - their purchasing power. This is related to the concern with functional or methodological equivalence discussed by cross-cultural researchers (Alasuutari et al., 2008), and it should give pause to any investigator engaged in a "direct replication" of a study. A replication needs to replicate the "deep," not the surface, structure of the original. It needs to replicate participants' construal from the original study, which, paradoxically, may require some reconfiguring of the original, particularly when construal processes in the new context obey a somewhat different logic from the original. Indeed, however faithful to the original on the surface, failure to reproduce the features that truly matter may cause

failures to replicate. For recent discussions on conceptual replication from psychology, see Monin et al. (2014).

Institutionalizing and scaling up experimental results

In fact, discerning how to construe the causal drivers of your effect for a replication presents similar challenges to those encountered when attempting to identify which factors should be "scaled-up." By scale-up, we mean either a large-scale replication of your experiment or the installation of your experimental manipulation as part of a public or private institution's regular operating procedure. Among the potential complications involved in scaling up an experimental manipulation is that the targeted population will most likely receive the intervention from a source that is different from that used in the original experimental evaluation. And that source (e.g., university, non-governmental organization or government) can matter a great deal for participants' construal of the intervention. Furthermore, the very fact that an intervention is no longer presented as a trial, or as merely "experimental," but, rather as an established policy, may itself generate significantly different construal.

To our knowledge, one of the most striking and sobering examples of a shift in participant's construal from an experimental to an institutionalized policy is the domestic violence experiment led by the National Institute of Justice (Garner et al., 1995). The experiment used an encouragement design for police officers responding to a call reporting a domestic incident. Officers were randomly assigned to arrest, mediate, or separate upon arrival at the scene through a color-coded notepad (though they could break with the randomization in the case of an emergency). The estimated effect of this experiment revealed the importance of arrests for preventing recidivism in domestic

abuse – arrests were found to reduce estimated future violence by more than 50%. The results were subsequently used to support laws promoting arrests of individuals believed to be responsible for spousal abuse. Follow-up estimates (Iyengar, 2010), however, found that these laws had *increased* the number of intimate partner homicides where they had been implemented.

Setting aside debates about the methods and findings from Iyengar (2010) vs. those from the National Institute of Justice experiments, we can ask how laws mandating arrest of abusive spouses could increase homicides. A plausible explanation boils down to violence victims' construal of a call to the police for help. During the National Institute of Justice's experiment, a call to the police was understood as just that – a call for help. The exact repercussions, what the police officer might do once on the scene, was uncertain. Clearly, abusive partners would never construe a call to the police as a welcome action; however, prior to laws mandating immediate arrest, these calls were not understood as requesting an arrest. Once inscribed into law, a call to the police meant a call to arrest the partner. Both partners in a domestic dispute presumably shared this new construal, which rendered it more consequential at least for the abuser, if not also for the abused. Certainty of arrest was a different construal from that which predominated the earlier "experimental" phases, and could explain why homicides rose following the introduction of the laws.

In sum, a target population's understanding of an intervention may change as the intervention scales up, comes from a different source, slightly changes form, or is no longer novel. Thinking about participants' construal in this way is also a means of understanding and anticipating negative externalities. As the scaled-up intervention misaligns with participants' construal, losses – fewer abused women saved, fewer plastic bags recycled – are thereby imposed on society at large.

5 Concluding thoughts

A fundamental tension in the behavioral sciences has long pitted the study of overt behavior, most blatantly represented by behaviorism, with that of covert mental processes, studied by the cognitive sciences. This tension is central to field experiments, where the ultimate goal is to change and measure actual behaviors, but where the design of the intervention rests heavily on participants' mental lives. In this chapter, we have focused on one fundamental aspect of mental life, namely construal.

Construal is how people come to represent everyday experiences. Some of it can be natural, immediate, and effortless, other parts can be conscious and effortful – the outcomes of both "System 1" and "System 2" thinking, respectively (Kahneman, 2011). This presents a significant challenge to researchers, because the resulting behavior, which ultimately is the thing of interest, will have been shaped by processes that are always difficult to observe, often hard to control, and ever-sensitive to minor nuance.

This chapter should have convinced you that field experiments are not off-the-shelf type instruments. They need to build shared construals in contexts where nuance really matters. Even in simple behavioral laboratory "games" that measure behavior in response to differing incentives, where moves and payoffs are all well defined, a mere alteration of the name of the game can significantly change participants' chosen strategies. In one study, participants (American college students as well as Israeli pilots) played an N-move Prisoner's Dilemma game, referred to as either the Wall Street or the Community game (Liberman et al., 2004). The results showed that labeling exerted far greater impact on the players' choice to cooperate versus defect – both in the first round and overall

– than anticipated by predictions of their peers based on the players' reputation.

Let's eat, Grandma!

Let's eat Grandma!

Small nuances can save lives. They can change strategic behaviors. And they can change the way that your experimental stimuli are construed during an experiment, or in an attempt at replication.

While we have no surefire method for managing construal, our advice is to think about and explore the various facets that might impact how your study might be construed. Rather than merely "delivering" the relevant information, think about the terms – community, wall street – used in the delivery, the context – church, school – in which it is being delivered, and who – woman, man, child – delivers the message as well as their potential role in this particular milieu. Similarly, when you attempt to replicate, worry about the participants' construal in the original study, not just the original stimuli or procedures. Repeat the psychologically important, not the superficial structure, of an experiment.

There is a famous anecdote about three baseball umpires talking about how they call balls and strikes. The first umpire calls them as he sees them, and the second umpire calls them as they are. The third umpire says they're nothing until he calls them. We think about construal that way. You might think participants simply construe based on what they see, or you might think they construe what's really there. But the fact is that there's nothing much happening in your study until participants have construed it. And the challenge is to handle that construal with great care.

References

Alasuutari, Pertti, Leonard Bickman, and Julia Brannen (2008), *The SAGE handbook of social research methods*. Sage.

Asch, Solomon E (1940), "Studies in the principles of judgments and attitudes: Ii. determination of judgments by group and by ego standards." *The Journal of Social Psychology*, 12, 433–465.

Asch, Solomon E (1948), "The doctrine of suggestion, prestige and imitation in social psychology." *Psychological review*, 55, 250.

Berger, Jonah, Marc Meredith, and S Christian Wheeler (2008), "Contextual priming: Where people vote affects how they vote." *Proceedings of the National Academy of Sciences*, 105, 8846–8849.

Bertrand, Marianne, Dean S Karlan, Sendhil Mullainathan, Eldar Shafir, and Jonathan Zinman (2010), "What's advertising content worth? evidence from a consumer credit marketing field experiment." *Quarterly Journal of Economics*, 125, 263–306.

Bruner, Jerome S (1957), "Going beyond the information given." *Contemporary approaches to cognition*, 1, 119–160.

Casey, Katherine, Rachel Glennerster, and Edward Miguel (2011), "Reshaping institutions: Evidence on aid impacts using a pre-analysis plan." Technical report, National Bureau of Economic Research.

Committee, The TOP Guidelines (2015), "Promoting an open research culture: The top guidelines for journals." *Working paper*, 1, 1–2.

Datta, Saugato and Sendhil Mullainathan (2014), "Behavioral design: A new approach to development policy." *Review of Income and Wealth*, 60, 7–35.

Eidelman, Scott and Christian S Crandall (2009), "A psychological advantage for the status quo." *Social and psychological bases of ideology and system justification*, 85–106.

Garner, Joel, Jeffrey Fagan, and Christopher Maxwell (1995), "Published findings from the spouse assault replication program: A critical review." *Journal of Quantitative Criminology*, 11, 3–28.

Gilovich, Thomas (1981), "Seeing the past in the present: The effect of associations to familiar events on judgments and decisions." *Journal of Personality and Social Psychology*, 40, 797.

Gilovich, Thomas, Dale Griffin, and Daniel Kahneman (2002), *Heuristics and biases: The psychology of intuitive judgment*. Cambridge University Press.

Gneezy, Uri and Aldo Rustichini (2000), "A fine is a price." *Journal of Legal Studies*, 29, 1–18.

Haushofer, Johannes and Jeremy Shapiro (2013), "The social costs of randomization."

Iyengar, Radha (2010), "Does arrest deter violence? comparing experimental and nonexperimental evidence on mandatory arrest laws." In *The Economics of Crime: Lessons For and From Latin America* (Rafael Di Tella, Sebastian Edwards, and Ernesto Schargrodsky, eds.), 421–452, NBER/University of Chicago Press.

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

Kahneman, Daniel (2013), "Foreword." In *The behavioral foundations of public policy* (Eldar Shafir, ed.), 7–9, Princeton University Press.

Kahneman, Daniel and Shane Frederick (2002), "Representativeness revisited: Attribute substitution in intuitive judgment." *Heuristics and biases: The psychology of intuitive judgment*, 49.

Kahneman, Daniel, Jack L Knetsch, and Richard H Thaler (1991), "Anomalies: The endowment effect, loss aversion, and status quo bias." *The journal of economic perspectives*, 193–206.

Kahneman, Daniel and Dale T Miller (1986), "Norm theory: Comparing reality to its alternatives." *Psychological review*, 93, 136–153.

Kelley, Harold H (1973), "The processes of causal attribution." *American psychologist*, 28, 107.

Lewin, Kurt and Martin Ed Gold (1999), *The complete social scientist: A Kurt Lewin reader.* American Psychological Association.

Liberman, Varda, Steven M Samuels, and Lee Ross (2004), "The name of the game: Predictive power of reputations versus situational labels in determining prisoner's dilemma game moves." *Personality and social psychology bulletin*, 30, 1175–1185.

Loftus, Geoffrey R and Norman H Mackworth (1978), "Cognitive determinants of fixation location during picture viewing." *Journal of Experimental Psychology: Human perception and performance*, 4, 565.

Markus, Hazel and Ziva Kunda (1986), "Stability and malleability of the self-concept." *Journal of personality and social psychology*, 51, 858.

Milgram, Stanley (1974), *Obedience to authority*.

Miller, Dale T and Deborah A Prentice (2013), "Psychological levers of behavior change." In *The behavioral foundations of public policy* (Eldar Shafir, ed.), 301–309, Princeton University Press.

Mischel, Walter, Ebbe B Ebbesen, and Antonette Raskoff Zeiss (1972), "Cognitive and attentional mechanisms in delay of gratification." *Journal of personality and social psychology*, 21, 204.

Monin, Benoît, Daniel M Oppenheimer, Melissa J Ferguson, Travis J Carter, Ran R Hassin, Richard J Crisp, Eleanor Miles, Shenel Husnu, Norbert Schwarz, Fritz Strack, et al. (2014), "Commentaries and rejoinder on klein et al.(2014)."

Morse, Stan and Kenneth J Gergen (1970), "Social comparison, self-consistency, and the concept of self." *Journal of personality and social psychology*, 16, 148.

Olken, Ben (2015), "Pre-analysis plans in economics." *Journal of Economic Perspectives*.

Orne, Martin T (1962), "On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications." *American psychologist*, 17, 776.

Paluck, Elizabeth Levy (2009), "Methods and ethics with research teams and ngos: Comparing experiences across the border of rwanda and democratic republic of congo." *Surviving Field Research: Working in Violent and Difficult Situations*, 38–56.

Redelmeier, Donald A, Jean-Pierre Molin, and Robert J Tibshirani (1995), "A randomised trial of compassionate care for the homeless in an emergency department." *The Lancet*, 345, 1131–1134.

Robins, Richard W, Holly M Hendin, and Kali H Trzesniewski (2001), "Measuring global self-esteem: Construct validation of a single-item measure and the rosenberg self-esteem scale." *Personality and social psychology bulletin*, 27, 151–161.

Rosnow, Ralph L and Robert Rosenthal (1997), *People studying people: Artifacts and ethics in behavioral research*. WH Freeman New York.

Ross, Lee and Richard E Nisbett (1991), *The person and the situation: Perspectives of social psychology.* Mcgraw-Hill Book Company.

Schank, Roger C and Robert P Abelson (2013), *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press.

Schwarz, Norbert and Jing Xu (2011), "Why don't we learn from poor choices? the consistency of expectation, choice, and memory clouds the lessons of experience." *Journal of Consumer Psychology*, 21, 142–145.

Seligman, Martin E (1970), "On the generality of the laws of learning." *Psychological review*, 77, 406.

Shafir, Eldar (2013), *The behavioral foundations of public policy*. Princeton University Press.

Skinner, Burrhus F (1960), "Pigeons in a pelican." *American Psychologist*, 15, 28.

Thaler, Richard H and Cass R Sunstein (2008), *Nudge*. Yale University Press.

Tversky, Amos and Daniel Kahneman (1973), "Availability: A heuristic for judging frequency and probability." *Cognitive psychology*, 5, 207–232.

Tversky, Amos and Daniel Kahneman (1981), "The framing of decisions and the psychology of choice." *Science*, 211, 453–458.

Vallone, Robert P, Lee Ross, and Mark R Lepper (1985), "The hostile media phenomenon: biased perception and perceptions of media bias in coverage of the beirut massacre." *Journal of personality and social psychology*, 49, 577.

Weber, Ernst Heinrich, Helen Elizabeth Ross, and David J Murray (1996), *EH Weber on the tactile senses*. Psychology Press.

Willis, Gordon B (2004), *Cognitive interviewing: A tool for improving questionnaire design*. Sage Publications.